

Florida State University Libraries

Electronic Theses, Treatises and Dissertations

The Graduate School

An Evaluation of Non-Gaussian Climate Statistics

Robert West

FLORIDA STATE UNIVERSITY
COLLEGE OF ARTS AND SCIENCES

AN EVALUATION OF NON-GAUSSIAN CLIMATE STATISTICS

By
ROBERT WEST

A Dissertation submitted to the
Department of Earth, Ocean and Atmospheric Science
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2020

Copyright © 2020 Robert West. All Rights Reserved.

Robert West defended this dissertation on 2 April 2020.
The members of the supervisory committee were:

Philip Sura
Professor Directing Dissertation

Bernd Berg
University Representative

Ming Cai
Committee Member

Allison Wing
Committee Member

Zhaohua Wu
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

For Mary Beth.

ACKNOWLEDGMENTS

I am most grateful to my major professor and mentor, Dr. Philip Sura, for his encouragement, wisdom and patience during my graduate studies at Florida State University. Thank you for generously sharing your advice and expertise. I will not forget our meetings and discussions, especially those over an espresso on a sunny afternoon.

I would like to thank Drs. Ming Cai, Zhaohua Wu, Allison Wing and Bernd Berg for serving as members of my committee. As an undergraduate student, Dr. Cai agreed to mentor me on my first research project and introduced me to the world of climate research. My first class as a graduate student was taught by Dr. Wu, who was kind enough to also serve on my Master's thesis committee. Dr. Wing was very generous to join my PhD committee and my research stands much improved due to her suggestions. Dr. Berg has been gracious enough to serve as the University Representative on my committee, even after his retirement.

Thank you also to Dr. Robert Hart, for his advice and always having a few minutes to talk, as well as to Dr. Robert Ellingson, who served on my Master's thesis committee.

Many thanks are due to lab colleagues Jonathan Christophersen and Maxime Perron, over which many years of discussion and collaboration has helped to shape the scientist I am today. I also thank my fellow undergraduate and graduate students, some of whom are still here and others who have moved on, as you've all had a part in my education.

Thanks to the HWind/RMS group for their support and encouragement to finish my degree.

Thank you to Kelly Hirai, who introduced me to Python and helped me become a better programmer. Thanks are also due to Shel McGuire, who helped me finish on time.

Thank you to my family, especially my parents, who have supported me throughout my education and provided me encouragement and love every step of the way. Dad, thank you for always making time to meet up, to discuss ideas or to read my manuscripts, even if it was late at night. Mom, thank you for all your hugs, words and notes of encouragement, especially when times weren't so good. I am grateful to Jim and Brenda, who generously and warmly welcomed me into their family. Thanks are due to my brother Tim, old friend and collaborator, who was always there to listen.

Last, but not least, thank you to Mary Beth. Without your love, support and encouragement, this dissertation would not have been finished. I love you everyday.

TABLE OF CONTENTS

List of Tables	vii
List of Figures	viii
Abstract	xvi
1 INTRODUCTION	1
1.1 Extreme events	1
1.2 A non-Gaussian atmosphere	2
1.3 Stochastic models of climate extremes	6
1.4 Roadmap	7
2 DATA	9
2.1 NOAA-CIRES-DOE Twentieth Century Reanalysis version 2c	9
2.2 500 hPa geopotential height	11
2.3 950 hPa air temperature	11
2.4 300 hPa relative vorticity	13
2.5 950 hPa zonal and meridional wind	13
2.6 500 hPa vertical velocity	15
3 SOME STOCHASTIC CLIMATE THEORY	19
3.1 Physical motivation	19
3.2 The stochastic differential equation	20
3.3 The Fokker-Planck equation	22
3.4 Stochastic dynamics of climate variability	24
3.5 Hasselmann’s climate model	26
3.6 Correlated additive and multiplicative noise model	27
4 THE STOCHASTICALLY GENERATED SKEWED DISTRIBUTION	34
4.1 Properties of the SGS distribution	34
4.2 Estimation of SGS parameters	36
4.3 The SGS Markov process	38
5 THE SGSDIST.JL PACKAGE	39
5.1 Installation	39
5.2 Creating an SGS distribution	40
5.3 Fitting an SGS distribution from data	40
5.4 Calculating the SGS pdf	41
5.5 Calculating the SGS CDF	42
5.6 Random number generation	43
5.7 Markov processes	45

6	STATISTICAL EVALUATION OF THE SGS DISTRIBUTION	49
6.1	Fitting an SGS distribution to a time series	49
6.2	Visual comparison with Q-Q plots	51
6.3	A goodness-of-fit test using the Kolmogorov-Smirnov statistic	53
6.4	Treating temporal correlations in reanalysis data	58
6.5	Results of the goodness-of-fit test	63
6.6	Model selection via likelihood ratio test	65
7	GLOBAL SGS EVALUATIONS OF 20CRV2C VARIABLES	68
7.1	950 hPa air temperature	70
7.2	500 hPa geopotential height	78
7.3	300 hPa relative vorticity	82
7.4	950 hPa zonal wind	87
7.5	950 hPa meridional wind	97
7.6	500 hPa vertical velocity	102
8	A COMPARISON OF REANALYSIS ERAS	111
8.1	Changes in non-Gaussianity from 1880-1947 to 1947-2014	111
8.2	Changes in the SGS distribution from 1880-1947 to 1947-2014	115
9	CONCLUDING DISCUSSION	124
9.1	Research overview and conclusions	124
9.2	Future work	129
	REFERENCES	132
	Biographical Sketch	137

LIST OF TABLES

6.1	Locations, higher moments and SGS parameters E , b and g of near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) 1947-2014 DJF 950 hPa air temperature anomaly time series.	51
6.2	Locations, lengths of time series (N) and pseudorandom time series (N') and lag time (τ) of near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) 1947-2014 DJF 950 hPa air temperature anomaly time series.	60
6.3	Locations and decorrelation times computed using the $1/e$ autocorrelation threshold, 0.1 autocorrelation threshold and the large-lag standard error (LLSE) at near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) 1947-2014 DJF 950 hPa air temperature anomaly time series.	62
6.4	The KS statistic and p-value computed from the SGS goodness-of-fit test for the near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) 1947-2014 DJF 950 hPa air temperature anomaly time series.	63
6.5	The KS statistic and p-value computed from the standard normal goodness-of-fit test for the near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) 1947-2014 DJF 950 hPa air temperature anomaly time series.	64
6.6	Likelihood ratio test statistic of the fitted SGS distribution against the standard normal distribution for the near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) 1947-2014 DJF 950 hPa air temperature anomaly time series.	66
7.1	Lookup table for the p-value significance thresholds, plausible (“good” fit) and implausible (“bad” fit) values and map colors for the goodness-of-fit (GOF) and the SGS-to-Gaussian comparison of the likelihood ratio test statistic (LRTS) statistical tests.	69
8.1	The KS statistic and p-value computed via goodness-of-fit test for the near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) DJF 950 hPa air temperature anomaly time series, comparing the estimated SGS distributions of DJF 1947-2014 to that of DJF 1880-1947.	117

LIST OF FIGURES

1.1	Maps of skewness and kurtosis of standardized DJF 500 hPa geopotential height anomalies from the NOAA-CIRES-DOE Twentieth Century Reanalysis Project version 2c dataset, 1851-2014.	3
2.1	Maps of skewness and kurtosis of standardized DJF 950 hPa air temperature anomalies from the 20CRv2c dataset, 1851-2014.	12
2.2	Maps of skewness and kurtosis of standardized DJF 300 hPa relative vorticity anomalies from the 20CRv2c dataset, 1851-2014.	14
2.3	Maps of skewness and kurtosis of standardized DJF 950 hPa zonal wind anomalies from the 20CRv2c dataset, 1851-2014.	16
2.4	Maps of skewness and kurtosis of standardized DJF 950 hPa meridional wind anomalies from the 20CRv2c dataset, 1851-2014.	17
2.5	Maps of skewness and kurtosis of standardized DJF 500 hPa vertical velocity anomalies from the 20CRv2c dataset, 1851-2014.	18
3.1	Example SGS pdf (red) with variance and skewness of one, kurtosis of five. The standard normal distribution (gray) is provided for reference.	29
3.2	Scatterplot of kurtosis versus skewness showing standardized 950 hPa air temperature anomalies (T') for wintertime (DJF) 20CRv2c data (1947-2014). The dark solid parabola indicates the skewness-kurtosis relationship $kurt \geq (\frac{3}{2}) skew^2$, where the light solid parabola is $kurt \geq (\frac{3}{2}) skew^2 - 0.972$. The dotted line represents the less stringent $kurt = skew^2 - 2$ constraint.	31
5.1	Plot of the example air temperature SGS pdf, showing the SGS pdf ($\sigma^2 = 1$, skew = 1, kurt = 5) (blue) in comparison with the step histogram of the sample air temperature data (red).	42
5.2	Plot of the "small E" SGS pdf, showing the SGS pdf (blue) in comparison with the standard normal pdf (red) when $E \rightarrow 0$	43
5.3	Plot of the SGS CDF (blue) obtained by estimating the SGS distribution from example air temperature data.	44
5.4	Plot of the CAM1D Markov process time series with default arguments ($dt=1/24$, $lambda=1$), where the statistics of the time series are informed from an SGS distribution with $\sigma^2 = 1$, skew = 1, kurt = 5.	45

5.5	Plot of the CAM1D Markov process time series with dt=1/12 , lambda=0.5 and a seed provided to the random number generator. The statistics of the time series are informed from an SGS distribution with $\sigma^2 = 1$, skew = 1, kurt = 5.	46
5.6	Step histogram (blue) of the previous CAM1D Markov process time series with dt=1/12 , lambda=0.5 along with the pdf (red) of the SGS distribution used to produce the Markov process time series.	47
5.7	Plot of the Hasselmann1D Markov process time series with default arguments (dt=1/24 , lambda=1), where the statistics of the time series are informed from a standard normal distribution.	48
6.1	Maps of skewness and kurtosis of standardized DJF 950 hPa 20CRv2c air temperature anomalies, 1947-2014. Point A (East Hawaii, 20°N, 138°W) and point B (Vancouver, 50°N, 124°W) are delineated on the map.	50
6.2	SGS pdfs (red) and stepwise histograms (blue) of the DJF 950 hPa 20CRv2c air temperature anomalies (1947-2014) located at near-Gaussian East Hawaii (top, 20°N, 138°W) and at non-Gaussian Vancouver (bottom, 50°N, 124°W). The standard normal distribution (gray) is provided for reference.	52
6.3	Q-Q plots of the DJF 950 hPa 20CRv2c air temperature anomaly (1947-2014) quantiles (blue dots) located at near-Gaussian East Hawaii (20°N, 138°W) compared against the quantiles of the standard normal distribution (top) and estimated SGS distribution (bottom). If the two distributions are similar, the air temperature quantiles will lie on the red $y = x$ line.	54
6.4	Q-Q plots of the DJF 950 hPa 20CRv2c air temperature anomaly (1947-2014) quantiles (blue dots) located at non-Gaussian Vancouver (50°N, 124°W) compared against the quantiles of the standard normal distribution (top) and estimated SGS distribution (bottom). If the two distributions are similar, the air temperature quantiles will lie on the red $y = x$ line.	55
6.5	Two example cumulative distribution functions (CDF or $F(x)$): an empirical cumulative distribution function (ECDF, blue) assembled from the example time series x and the estimated SGS cumulative distribution function (red) obtained by fitting an SGS distribution to x . The Kolmogorov-Smirnov statistic, D , is the maximum difference between the two CDFs, shown by the black arrow.	56
6.6	ECDF (blue) of the DJF 950 hPa 20CRv2c air temperature anomalies (1947-2014) and estimated SGS CDF (red) located at near-Gaussian East Hawaii (top, 20°N, 138°W) and Vancouver (bottom, 50°N, 124°W) compared against ECDFs obtained from pseudorandom time series scaled by decorrelation times with various cutoff thresholds ($1/e$, 0.1, large-lag standard error or LLSE). The equivalent sample size (N') for each ECDF is included.	61

7.1	Maps of SGS parameters E (top), b (middle) and g (bottom) estimated from standardized DJF 950 hPa 20CRv2c air temperature anomalies, 1947-2014.	71
7.2	Map of the SGS model violations encountered when fitting an SGS distribution to time series of standardized DJF 950 hPa 20CRv2c air temperature anomalies, 1947-2014. Violations $E^2 < 0$ (cyan) and $b^2 < 0$ (olive) are most prominent.	72
7.3	Decorrelation time scale in days computed using the $1/e$ autocorrelation threshold for the standardized DJF 950 hPa 20CRv2c air temperature anomalies, 1947-2014.	72
7.4	p-values for the estimated SGS (top) and standard normal (bottom) goodness-of-fit (GOF) test computed using 100 bootstrap iterations for all gridded time series of standardized DJF 950 hPa 20CRv2c air temperature anomalies, 1947-2014. Gray points correspond to a plausible SGS fit, while points gridded blue, especially those with values less than 0.1, correspond to an implausible SGS fit.	74
7.5	Bootstrap CDFs (olive) generated during the standard normal goodness-of-fit (GOF) test for an equatorial time series of 950 hPa 20CRv2c air temperature anomalies, 1947-2014, with an equivalent sample size (N') of 112. The estimated SGS CDF (red) is compared with the ECDF of the time series of anomalies (blue). The standard normal CDF (gray) is provided for reference.	75
7.6	Skewness-kurtosis scatter plot showing the goodness-of-fit (GOF) p-value relative to the sample skewness and kurtosis computed for each time series of 950 hPa 20CRv2c air temperature anomalies, 1947-2014. The solid black parabola indicates the skewness-kurtosis relationship $kurt \geq (\frac{3}{2}) skew^2$, where the light gray solid parabola is $kurt \geq (\frac{3}{2}) skew^2 - r$ where r is empirically fit to the data. The dotted gray line represents the $kurt = skew^2 - 2$ constraint.	76
7.7	p-values of the likelihood ratio test statistic (LRTS) comparison between the estimated SGS and standard normal distributions for all gridded time series of standardized DJF 950 hPa 20CRv2c air temperature anomalies, 1947-2014. Gray points correspond to grid points where the SGS distribution is statistically preferred to the standard normal distribution. Points gridded blue, especially those with values much greater than 0.1, correspond to a statistically insignificant preference for the SGS fit or a preference for the standard normal distribution.	77
7.8	Maps of skewness and kurtosis of standardized DJF 500 hPa geopotential height anomalies from the 20CRv2c dataset, 1947-2014.	79
7.9	Map of the SGS model violations encountered when fitting an SGS distribution to time series of standardized DJF 500 hPa 20CRv2c geopotential height anomalies, 1947-2014. Violations $E^2 < 0$ (cyan) and $b^2 < 0$ (olive) are most prominent.	80
7.10	Decorrelation time scale in days computed using the $1/e$ autocorrelation threshold for the standardized DJF 500 hPa 20CRv2c geopotential height anomalies, 1947-2014.	80

7.11	p-values for the estimated SGS (top) and standard normal (bottom) goodness-of-fit (GOF) test computed using 100 bootstrap iterations for all gridded time series of standardized DJF 500 hPa 20CRv2c geopotential height anomalies, 1947-2014. Gray points correspond to a plausible SGS fit, while points gridded blue, especially those with values less than 0.1, correspond to an implausible SGS fit.	81
7.12	Skewness-kurtosis scatter plot showing the goodness-of-fit (GOF) p-value relative to the sample skewness and kurtosis computed for each time series of 500 hPa 20CRv2c geopotential height anomalies, 1947-2014. The solid black parabola indicates the skewness-kurtosis relationship $kurt \geq (\frac{3}{2}) skew^2$, where the light gray solid parabola is $kurt \geq (\frac{3}{2}) skew^2 - r$ where r is empirically fit to the data. The dotted gray line represents the $kurt = skew^2 - 2$ constraint.	83
7.13	p-values of the likelihood ratio test statistic (LRTS) comparison between the estimated SGS and standard normal distributions for all gridded time series of standardized DJF 500 hPa 20CRv2c geopotential height anomalies, 1947-2014. Gray points correspond to grid points where the SGS distribution is statistically preferred to the standard normal distribution. Points gridded blue, especially those with values much greater than 0.1, correspond to a statistically insignificant preference for the SGS fit or a preference for the standard normal distribution.	84
7.14	Maps of skewness and kurtosis of standardized DJF 300 hPa relative vorticity anomalies from the 20CRv2c dataset, 1947-2014.	85
7.15	Map of the SGS model violations encountered when fitting an SGS distribution to time series of standardized DJF 300 hPa 20CRv2c relative vorticity anomalies, 1947-2014. Violations $E^2 < 0$ (cyan) and $b^2 < 0$ (olive) are most prominent.	86
7.16	Decorrelation time scale in days computed using the $1/e$ autocorrelation threshold for the standardized DJF 300 hPa 20CRv2c relative vorticity anomalies, 1947-2014. . . .	87
7.17	p-values for the estimated SGS (top) and standard normal (bottom) goodness-of-fit (GOF) test computed using 100 bootstrap iterations for all gridded time series of standardized DJF 300 hPa 20CRv2c relative vorticity anomalies, 1947-2014. Gray points correspond to a plausible SGS fit, while points gridded blue, especially those with values less than 0.1, correspond to an implausible SGS fit.	88
7.18	Skewness-kurtosis scatter plot showing the goodness-of-fit (GOF) p-value relative to the sample skewness and kurtosis computed for each time series of 300 hPa 20CRv2c relative vorticity anomalies, 1947-2014. The solid black parabola indicates the skewness-kurtosis relationship $kurt \geq (\frac{3}{2}) skew^2$, where the light gray solid parabola is $kurt \geq (\frac{3}{2}) skew^2 - r$ where r is empirically fit to the data. The dotted gray line represents the $kurt = skew^2 - 2$ constraint.	89
7.19	p-values of the likelihood ratio test statistic (LRTS) comparison between the estimated SGS and standard normal distributions for all gridded time series of standardized DJF 300 hPa 20CRv2c relative vorticity anomalies, 1947-2014. Gray points correspond to	

	grid points where the SGS distribution is statistically preferred to the standard normal distribution. Points gridded blue, especially those with values much greater than 0.1, correspond to a statistically insignificant preference for the SGS fit or a preference for the standard normal distribution.	90
7.20	Maps of skewness and kurtosis of standardized DJF 950 hPa zonal wind anomalies from the 20CRv2c dataset, 1947-2014.	91
7.21	Map of the SGS model violations encountered when fitting an SGS distribution to time series of standardized DJF 950 hPa 20CRv2c zonal wind anomalies, 1947-2014. Violations $E^2 < 0$ (cyan) and $b^2 < 0$ (olive) are most prominent.	92
7.22	Decorrelation time scale in days computed using the $1/e$ autocorrelation threshold for the standardized DJF 950 hPa 20CRv2c zonal wind anomalies, 1947-2014.	93
7.23	p-values for the estimated SGS (top) and standard normal (bottom) goodness-of-fit (GOF) test computed using 100 bootstrap iterations for all gridded time series of standardized DJF 950 hPa 20CRv2c zonal wind anomalies, 1947-2014. Gray points correspond to a plausible SGS fit, while points gridded blue, especially those with values less than 0.1, correspond to an implausible SGS fit.	94
7.24	Skewness-kurtosis scatter plot showing the goodness-of-fit (GOF) p-value relative to the sample skewness and kurtosis computed for each time series of 950 hPa 20CRv2c zonal wind anomalies, 1947-2014. The solid black parabola indicates the skewness-kurtosis relationship $kurt \geq (\frac{3}{2}) skew^2$, where the light gray solid parabola is $kurt \geq (\frac{3}{2}) skew^2 - r$ where r is empirically fit to the data. The dotted gray line represents the $kurt = skew^2 - 2$ constraint.	95
7.25	p-values of the likelihood ratio test statistic (LRTS) comparison between the estimated SGS and standard normal distributions for all gridded time series of standardized DJF 950 hPa 20CRv2c zonal wind anomalies, 1947-2014. Gray points correspond to grid points where the SGS distribution is statistically preferred to the standard normal distribution. Points gridded blue, especially those with values much greater than 0.1, correspond to a statistically insignificant preference for the SGS fit or a preference for the standard normal distribution.	96
7.26	Maps of skewness and kurtosis of standardized DJF 950 hPa meridional wind anomalies from the 20CRv2c dataset, 1947-2014.	98
7.27	Map of the SGS model violations encountered when fitting an SGS distribution to time series of standardized DJF 950 hPa 20CRv2c meridional wind anomalies, 1947-2014. Violations $E^2 < 0$ (cyan) and $b^2 < 0$ (olive) are most prominent.	99
7.28	Decorrelation time scale in days computed using the $1/e$ autocorrelation threshold for the standardized DJF 950 hPa 20CRv2c meridional wind anomalies, 1947-2014.	99

7.29	p-values for the estimated SGS (top) and standard normal (bottom) goodness-of-fit (GOF) test computed using 100 bootstrap iterations for all gridded time series of standardized DJF 950 hPa 20CRv2c meridional wind anomalies, 1947-2014. Gray points correspond to a plausible SGS fit, while points gridded blue, especially those with values less than 0.1, correspond to an implausible SGS fit.	100
7.30	Skewness-kurtosis scatter plot showing the goodness-of-fit (GOF) p-value relative to the sample skewness and kurtosis computed for each time series of 950 hPa 20CRv2c meridional wind anomalies, 1947-2014. The solid black parabola indicates the skewness-kurtosis relationship $kurt \geq (\frac{3}{2}) skew^2$, where the light gray solid parabola is $kurt \geq (\frac{3}{2}) skew^2 - r$ where r is empirically fit to the data. The dotted gray line represents the $kurt = skew^2 - 2$ constraint.	101
7.31	p-values of the likelihood ratio test statistic (LRTS) comparison between the estimated SGS and standard normal distributions for all gridded time series of standardized DJF 950 hPa 20CRv2c meridional wind anomalies, 1947-2014. Gray points correspond to grid points where the SGS distribution is statistically preferred to the standard normal distribution. Points gridded blue, especially those with values much greater than 0.1, correspond to a statistically insignificant preference for the SGS fit or a preference for the standard normal distribution.	102
7.32	Maps of skewness and kurtosis of standardized DJF 500 hPa vertical velocity anomalies from the 20CRv2c dataset, 1947-2014.	104
7.33	Map of the SGS model violations encountered when fitting an SGS distribution to time series of standardized DJF 500 hPa 20CRv2c vertical velocity anomalies, 1947-2014. Violations $E^2 < 0$ (cyan), $b^2 < 0$ (olive) and $ skew > skew_{max}$ (purple) are prominent.	105
7.34	Decorrelation time scale in days computed using the $1/e$ autocorrelation threshold for the standardized DJF 500 hPa 20CRv2c vertical velocity anomalies, 1947-2014.	105
7.35	p-values for the estimated SGS (top) and standard normal (bottom) goodness-of-fit (GOF) test computed using 100 bootstrap iterations for all gridded time series of standardized DJF 500 hPa 20CRv2c vertical velocity anomalies, 1947-2014. Gray points correspond to a plausible SGS fit, while points gridded blue, especially those with values less than 0.1, correspond to an implausible SGS fit.	106
7.36	SGS pdf (red) and stepwise histograms (blue) of the DJF 500 hPa 20CRv2c vertical velocity anomalies (1947-2014) located at $20^\circ N$, $150^\circ W$. The standard normal distribution (gray) is provided for reference.	107
7.37	Bootstrap CDFs (olive) generated during the standard normal goodness-of-fit (GOF) test for DJF 500 hPa 20CRv2c vertical velocity anomalies (1947-2014) located at $20^\circ N$, $150^\circ W$, with an equivalent sample size (N') of 2046. The estimated SGS CDF (red) is compared with the ECDF of the time series of anomalies (blue). The standard normal CDF (gray) is provided for reference.	108

7.38	Skewness-kurtosis scatter plot showing the goodness-of-fit (GOF) p-value relative to the sample skewness and kurtosis computed for each time series of 500 hPa 20CRv2c vertical velocity anomalies, 1947-2014. The solid black parabola indicates the skewness-kurtosis relationship $kurt \geq \left(\frac{3}{2}\right) skew^2$, where the light gray solid parabola is $kurt \geq \left(\frac{3}{2}\right) skew^2 - r$ where r is empirically fit to the data. The dotted gray line represents the $kurt = skew^2 - 2$ constraint.	109
7.39	p-values of the likelihood ratio test statistic (LRTS) comparison between the estimated SGS and standard normal distributions for all gridded time series of standardized DJF 500 hPa 20CRv2c vertical velocity anomalies, 1947-2014. Gray points correspond to grid points where the SGS distribution is statistically preferred to the standard normal distribution. Points gridded blue, especially those with values much greater than 0.1, correspond to a statistically insignificant preference for the SGS fit or a preference for the standard normal distribution.	110
8.1	Maps of the difference in skewness and kurtosis of standardized 20CRv2c DJF 950 hPa air temperature anomalies from the historical era (1880-1947) to the modern era (1947-2014).	112
8.2	Maps of the difference in skewness and kurtosis of standardized 20CRv2c DJF 500 hPa geopotential height anomalies from the historical era (1880-1947) to the modern era (1947-2014).	113
8.3	Maps of the difference in skewness and kurtosis of standardized 20CRv2c DJF 300 hPa relative vorticity anomalies from the historical era (1880-1947) to the modern era (1947-2014).	114
8.4	SGS pdfs (red, purple) and stepwise histograms (blue, cyan) of the 20CRv2c DJF 950 hPa air temperature anomalies from the historical (1880-1947) and modern (1947-2014) eras at near-Gaussian East Hawaii (top) and non-Gaussian Vancouver (bottom), showing little change between the two time periods. The standard normal distribution (gray) is provided for reference.	116
8.5	ECDFs (blue, purple) of the DJF 950 hPa 20CRv2c air temperature anomalies and estimated SGS CDFs (red, purple) from the historical (1880-1947) and modern (1947-2014) eras at at near-Gaussian East Hawaii (top) and Vancouver (bottom). A confidence envelope created with ECDFs obtained from bootstrapping 100 pseudorandom time series is included for comparison.	118
8.6	p-values obtained from the goodness-of-fit (GOF) test comparing changes in SGS distributions in the modern era (1947-2014) with the historical (1880-1947) era. Gray points correspond to a plausible difference in the SGS distribution between eras, while blue points signify SGS distributions that are not statistically different across the two eras.	120
8.7	p-values obtained from the likelihood ratio test statistic (LRTS) comparing changes in SGS distributions in the modern era (1947-2014) with the historical (1880-1947) era.	

Gray points correspond to a plausible difference in the SGS distribution between eras,
while blue points signify SGS distributions that are not statistically different across
the two eras. 123

ABSTRACT

Earth’s atmosphere-ocean system is distinguished by its variability over a wide range of time scales. The non-linear interactions between these time scales are complex and are further complicated by the large number of subsystems and modes in the atmosphere-ocean system. Here, we explore a stochastic model developed by Sardeshmukh and Sura which uses correlated additive and multiplicative (CAM) noise and relies on a state-dependent (multiplicative) noise forcing to represent the multi-scale interactions between weather and climate.

An important problem in climate variability is the statistical representation of extreme weather and climate events. While a description of the tails of a probability density function (pdf) is essential for modeling extreme events, an understanding of the full pdf is required to capture the full dynamics of the atmosphere-ocean system. On daily scales, the statistics of the large-scale atmospheric circulation are non-Gaussian. A one-dimensional pdf produced by the CAM noise model, or stochastically generated skewed (SGS) distribution, attempts to probabilistically represent the non-Gaussian statistics of atmospheric climate anomalies.

This study evaluates the ability of the SGS distribution to represent the non-Gaussian statistics of several atmospheric variables using NOAA-CIRES-DOE Twentieth Century Reanalysis Project version 2c (20CRv2c) dataset. A method of moments SGS parameter estimation technique described Sardeshmukh et al (2015) is implemented in a Julia software package and applied to global gridded time series of reanalysis data. Goodness-of-fit tests show the SGS distribution performs well in regions of near-zero and positive kurtosis, but produces statistically implausible with time series with negative sample kurtosis. However, the SGS distribution is found to outperform the standard normal (Gaussian) distribution at nearly all gridded locations, even where the SGS fit is poor. The SGS distributions of two 67 year 20CRv2c periods are also compared, where few significant changes in the shape of the SGS distribution are found.

CHAPTER 1

INTRODUCTION

The stochastically generated skewed (SGS) distribution is a dynamically consistent stochastic climate model that describes the non-Gaussian statistics of daily atmospheric data.

1.1 Extreme events

Extreme events (such as major hurricanes or severe earthquakes) typically impose a disproportionately large impact on lives and infrastructure, despite their infrequent occurrence. Extreme events are found within nature, technology and society and their incidence can be subject to a measure of subjectivity (Albeverio et al. (2006)). In order to properly study the occurrence and impact of extreme weather and climate events, about which this research is concerned, a more rigorous definition is needed.

Extreme events may be thought of as difficult-to-predict phenomena outside of normal expectations, where *normal* corresponds to the Gaussian distribution. One may also visualize the tails of a probability distribution function (pdf) as the statistical regions where extreme events occur. If an extreme event is beyond normal expectations, then the event will fall outside the tails of the Gaussian pdf i.e., the extreme event will be non-Gaussian (AghaKouchak et al. (2012)).

As such, understanding extreme weather and climate events requires the probabilistic representation of the tails of a non-Gaussian pdf. This statistical representation is crucial in the development of sophisticated and efficient risk and disaster management systems, especially within strategies to adapt to changes in the frequency and magnitude of extreme weather associated with anthropogenic climate change. It is important to note that understanding of the tails of the pdf is only part of the overall goal in the analysis of extreme events – understanding of the *full* pdf is to statistically encapsulate the full dynamics of the system. This knowledge correspondingly stimulates a deeper understanding of climate variability as a whole.

1.2 A non-Gaussian atmosphere

The statistics of many daily atmospheric variables (such as air temperature, geopotential height, zonal and meridional wind) are non-Gaussian, which has major implications for the understanding of climate statistics and the study of extreme events.¹ To relate the notion of Gaussianity with a probability distribution, the statistical moments used to describe the shape of a pdf first must be defined. If X is a random variable with an expected value of $\mu = E[X]$, the n^{th} central moment, $\langle x^n \rangle$, may be written as

$$\langle x^n \rangle = E[(x - \mu)^n] . \quad (1.1)$$

Using Equation (1.1), the first two central moments – mean and variance – may be written as

$$\begin{aligned} \langle x^1 \rangle &= \mu = E[(x - \mu)] \\ \langle x^2 \rangle &= \sigma^2 = E[(x - \mu)^2] . \end{aligned} \quad (1.2)$$

Skewness and kurtosis are defined as the third and fourth standardized moments, which normalize the third and fourth central moments by σ^n :

$$\begin{aligned} \langle x^3 \rangle &= \text{skew} = \frac{E[(x - \mu)^3]}{\sigma^3} \\ \langle x^4 \rangle &= \text{kurt} = \frac{E[(x - \mu)^4]}{\sigma^4} . \end{aligned} \quad (1.3)$$

For reference, a standard normal (Gaussian) distribution has zero mean, unit variance and a skewness and excess kurtosis of zero. As the kurtosis of a univariate normal distribution is equal to 3, the excess kurtosis is defined as the kurtosis minus 3. However, for brevity the excess kurtosis will be referred to hereafter simply as “kurtosis.”

The non-Gaussianity of the atmosphere can be demonstrated by composing maps of skewness and kurtosis from daily atmospheric data sets² of standardized anomalies. Figure 1.1 shows the skewness and kurtosis of the standardized anomalies of the 500 hPa geopotential height field from the NOAA-CIRES-DOE Twentieth Century Reanalysis Project version 2c dataset (Compo et al. (2011)).³ If the atmosphere were Gaussian, the skewness and kurtosis in Figure 1.1 would be zero.

¹Most observations obtained from nature are typically non-Gaussian in some fashion, unless they are excessively averaged and become subjected to the central limit theorem (AghaKouchak et al. (2012)).

²More detailed discussion on the dataset used in this work is found in Chapter 2. For now, the reanalysis data here is being used to highlight the non-Gaussian statistics of daily atmospheric data.

³For a more extensive analysis of the climatology of non-Gaussianity, see Perron and Sura (2013).

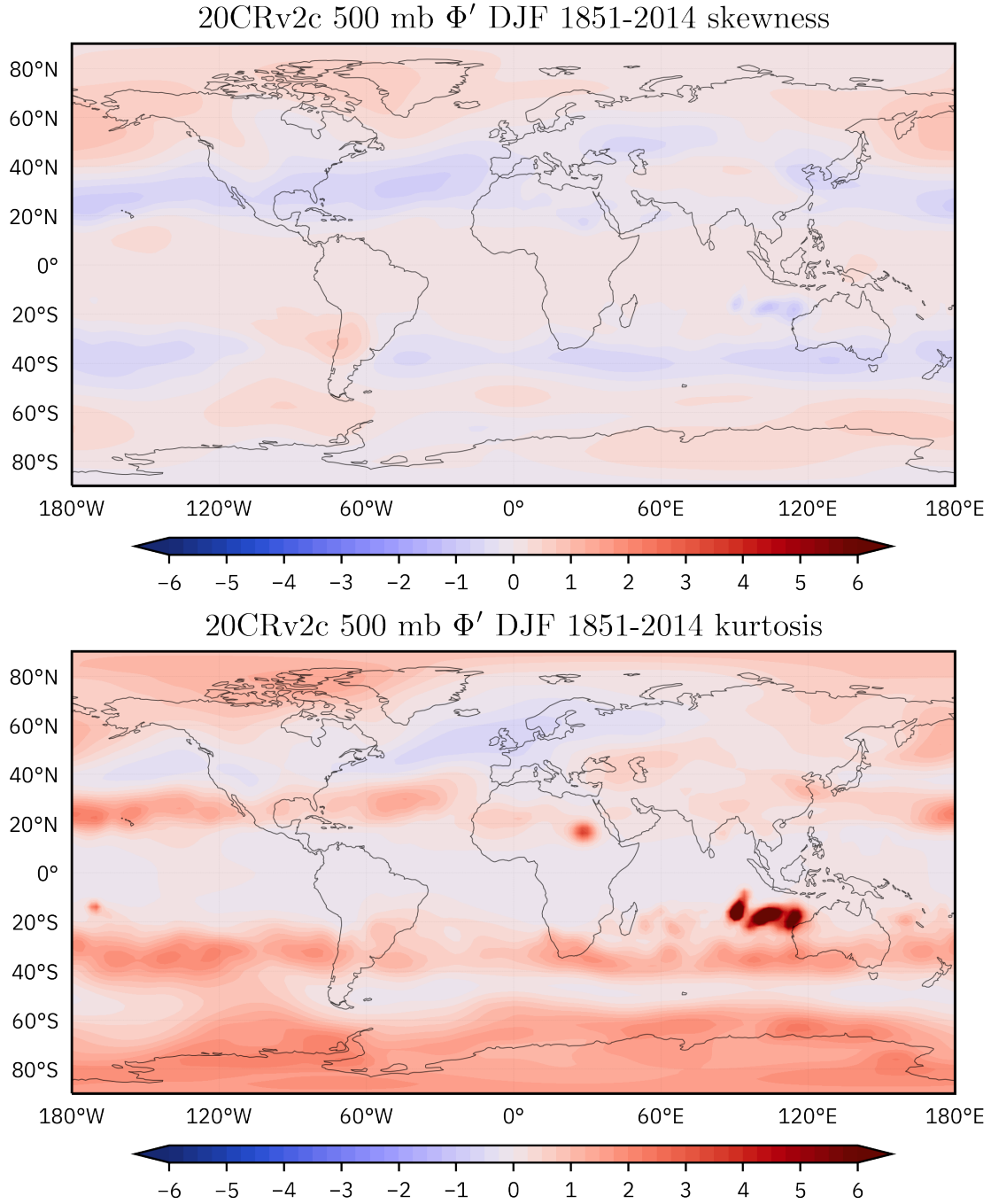


Figure 1.1: Maps of skewness and kurtosis of standardized DJF 500 hPa geopotential height anomalies from the NOAA-CIRES-DOE Twentieth Century Reanalysis Project version 2c dataset, 1851-2014.

Standardized anomalies are often used as the variable of study in climate statistics, as they provide a normalized, dimensionless unit of measurement that can be compared across different climate variables and probability distributions. Here, when calculating the standardized anomalies from a field such as a daily-averaged reanalysis dataset, one first calculates the mean annual cycle for each day of the calendar year at each grid point. The annual cycle is then subtracted from the field to produce the anomalies for each day of the year at each location. The anomalies are then divided by the standard deviation of each location, which yields a time series of standardized anomalies at each grid location. Further processing may be done when seasonal (DJF, MAM, JJA, SON) or geographical selection of data is desired.

The global skewness and kurtosis patterns for the DJF 500 hPa geopotential height anomalies in Figure 1.1 shows a banding structure with respect to latitude. Both the polar and equatorial regions are positively skewed, while the mid-latitudes of both hemispheres are negatively skewed. The skewness minima and maxima are located over the oceans, with the area of largest magnitude positioned over the North Atlantic. The skewness map also appears to show the Aleutian Low, a large scale circulation pattern that typically forms in the northern Pacific Ocean, as a region of positive skewness. There also seems to be a tilt of larger skewness values toward the Northern Hemisphere, which may be expected due to the winter months being more synoptically active (Perron and Sura (2013)).

The kurtosis map only partly resembles the distribution of skewness values, as the Northern Hemisphere polar regions, save the western coast of Greenland, have negative kurtosis. The banding structure in the Southern Hemisphere is more prevalent, with alternating regions of positive and negative kurtosis extending from the equator to the pole.

The patterns of non-Gaussianity in geopotential height fields such as Figure 1.1 have been studied previously by White (1980), Trenberth and Mo (1985) and Nakamura and Wallace (1991), which relate the spatial structure in geopotential height skewness and kurtosis to the locations of storm tracks and seasonal variations of troughs and ridges. Rennert and Wallace (2009) found that the cross-frequency coupling of the low ($< 30 \text{ days}^{-1}$) and medium ($6 - 30 \text{ days}^{-1}$) frequencies of the variability of 500 hPa geopotential height field contributes significantly to the observed skewness for the northern Atlantic and Pacific. Luxford and Woollings (2012) have demonstrated that the skewness of geopotential height and zonal wind fields are linked with the locations of jet streams.

While the non-Gaussianity of the climate variables used in this study are discussed in Chapter 2, further discussion of the physical mechanisms linked to extreme events and non-Gaussian statistics is provided by Hoskins and Woollings (2015).

A few studies on the non-Gaussianity of reanalysis data sets have been produced. Perron and Sura (2013) provide a thorough climatology for several variables of daily atmospheric data from the NCEP/NCAR Reanalysis 1 dataset⁴, including vertical cross-sections of zonally averaged variables that also show significant non-Gaussianity signatures. Petoukhov et al. (2008) found a significant deviation from Gaussianity of the skewness and mixed third-order moments for several atmospheric variables obtained from ERA-40 reanalysis data.⁵

The relationship of non-Gaussian extreme events with anthropogenic climate change (ACC) has also been investigated. Sardeshmukh et al. (2015) show that some extreme climate variability can be accounted for by simply considering pdfs in a non-Gaussian framework, where deviations from Gaussianity in the pdf tail can result in a modification of extreme event probabilities by an order of magnitude or more. This result has implications for estimating any change in the frequency of extreme events that may be attributed to ACC – models that assume a Gaussian pdf shape risk misattribution of extreme events to changes in the mean or higher moments rather than the (potentially) heavier tails of a non-Gaussian pdf. Indeed, Lopez et al. (2018) finds that the future emergence of regional heat waves attributed to anthropogenic forcing in the western and Great Lakes regions of the United States are witnessed in changes in the higher moments, not just the mean surface temperature. Conversely, the emergence of regional heat waves in the Great Plains attributed to ACC are found to be primarily driven by enhanced internal variability, not a change in the asymmetry of the pdf or heaviness of the pdf tails.

The physical basis underlying the dynamics of the non-Gaussian shape of the pdf and the observed extreme events is complex and still not well understood. However, models motivated by stochastic climate theory have been put forward as a means to reproduce and gain insight into the non-Gaussian statistics of daily climate data.

⁴See Kalnay et al. (1996) for more information on the NCEP/NCAR Reanalysis 1 dataset.

⁵For more on the ERA-40 dataset, see Uppala et al. (2005).

1.3 Stochastic models of climate extremes

The use of stochastic models to study climate variability and extreme events arose out of the idea to represent the natural variability of physical timescales as a coupling of slowly changing systems to rapidly varying physical processes (Sura and Sardeshmukh (2008)). In this sense, weather and climate may be defined as the manifestation of rapid and gradual variations in geophysical phenomena, respectively. Though weather and climate can be defined as systems spanning fast-slow time scales, it is also possible to analytically calculate the length of time scales in order to compare their natural variability e.g., Lovejoy (2013). In either case the underlying timescale separation is a simplification of the atmospheric system that enables a stochastic model to represent the dynamics of the timescale with noise.

Hasselmann (1976) was the first to invoke stochastic differential equations to approximate climate as a slowly decorrelating process and weather as a rapidly decorrelating phenomena, where climate is randomly “forced” by fluctuations in the weather approximated by white noise. While Hasselmann’s model produces the familiar red-noise climate spectrum, it is unable to reproduce the non-Gaussian statistics featured in observations and linked to extreme climate events (Sura (2011)).

To account for the presence of observed non-Gaussian statistics in atmosphere-ocean data, Sura and Sardeshmukh (2008) and Sardeshmukh and Sura (2009) introduced a stochastic climate model featuring correlated additive and multiplicative (CAM) noise that reproduces observed non-Gaussian skewness and kurtosis patterns while retaining the red climate spectrum of Hasselmann (1976). Sura and Sardeshmukh (2008) show that solving for kurtosis as a function of skewness results in a parabolic relationship that persists throughout many atmospheric-oceanic systems. Additionally, a solution of the CAM noise stochastic model described by Sardeshmukh and Sura (2009) indicates that the pdf tails of observed anomalies follow power law distributions. This behavior was first examined using log-log plots of observed and model-derived geopotential height and vorticity data in Sardeshmukh and Sura (2009) and Sura and Perron (2010) as well as in situ SST data in Sura (2010) and daily-averaged reanalysis data in West (2012). More on the Hasselmann and CAM noise stochastic models will be presented later in Chapter 3 and Chapter 4.

A major result obtained by Sardeshmukh and Sura (2009) is that the full pdf of the stochastic CAM noise model may be found analytically in the one-dimensional system. This pdf was called the stochastically generated skewed (SGS) distribution by Sardeshmukh and Sura (2009). Sardeshmukh

et al. (2015) further introduces a method of moments to estimate the parameters of the SGS pdf from the statistics of empirical climate data. Moreover, a Markov process model is derived by Sardeshmukh and Sura (2009), which has great utility for generating time series of statistically equivalent SGS distributions. This capability of the SGS distribution is a useful tool for sampling the tail of the non-Gaussian pdf, where large samples of extreme events can be generated with low computational expense.

A more thorough review of the history and scope of stochastic climate theory is found in Franzke et al. (2015) and Sura and Hannachi (2015). Further background on stochastic climate theory and the relationship with non-Gaussian statistics is also developed in Chapter 3.

1.4 Roadmap

Proceedings from the *Large-Scale Atmospheric Controls of Extreme Weather Events and Novel Predictability Pathways* workshop hosted at Stockholm University in October 2017 (summarized by Messori et al. (2018)) focused on the development of techniques to analyze extreme events and their relationship to large-scale atmospheric flows. Specifically, the problem of sampling large numbers of extreme events was discussed in the context of studying extreme heat waves from climate model output. This discussion highlighted a need for long integrations that encapsulate a large sample of extreme events (whatever they may be) so that the dynamics of the system that creates the extreme events may be examined.

To begin addressing such a requirement, this research will evaluate how well the SGS distribution is able to represent the non-Gaussian climate statistics of daily-averaged reanalysis data, specifically those of the NOAA-CIRES-DOE Twentieth Century Reanalysis Project version 2c (20CRv2c) dataset (Compo et al. (2011)) discussed in Chapter 2. Chapter 3 discusses stochastic climate theory in more detail including the origins of non-Gaussian climate theory via Hasselmann (1976) and the properties of the CAM noise stochastic model developed by Sura and Sardeshmukh (2008) and Sardeshmukh and Sura (2009). Chapter 4 focuses on the form and characteristics of the SGS distribution, including the distribution moments, parameters and the methodology to fit an SGS distribution to a time series of data.

To evaluate the SGS distribution with respect to reanalysis data, numerical routines to create, extend and evaluate SGS distributions have been developed and organized as a Julia⁶ module. Described in Chapter 5, this module, `SGSDist.jl`, facilitates the:

- Creation of an SGS distribution by fitting distribution parameters from time series data
- Calculation of the SGS pdf, CDF and other statistical quantities
- Drawing of random numbers from an SGS distribution
- Creation of synthetic time series using a Markov process with the time series statistics informed from a specified SGS distribution

In Chapter 6, the SGS distribution is estimated at two locations, which serve as two case studies for describing the statistical evaluation methodology. Chapter 7 expands the evaluation of the SGS distribution to a global scale by testing the ability of the distribution to reproduce the persistent wintertime statistics of several atmospheric variables from the 20CRv2c dataset over years 1947-2014. Statistical comparisons to the Gaussian distribution are also made using a model selection test. Chapter 8 compares the statistical results of the 1947-2014 era to the early 20CRv2c record (1880-1947) and investigates changes in the shape of the SGS distribution, which has consequences for some underlying assumptions used in the CAM noise model. Discussion and conclusions drawn from these results are presented in Chapter 9.

⁶The Julia language, created in 2012, is a dynamically-typed language similar to Python with the performance of Fortran or C. It was developed for use as a scientific programming language, in the mode of MATLAB or Mathematica. More information on Julia can be found in Bezanson et al. (2017) or at the project’s documentation page, <https://docs.julialang.org/>.

CHAPTER 2

DATA

The following chapter outlines the reanalysis dataset that will be used for the fitting and evaluating of SGS distributions. The specific variables and levels chosen for analysis are identified, along with a brief discussion of their large-scale non-Gaussian statistics.

2.1 NOAA-CIRES-DOE Twentieth Century Reanalysis version 2c

This research uses the NOAA-CIRES-DOE Twentieth Century Reanalysis Project version 2c (hereafter 20CRv2c) dataset for daily observations of several atmospheric variables. The 20CRv2c dataset is an updated release of the NOAA-CIRES-DOE 20th Century Reanalysis (20CRv2) dataset (see Compo et al. (2011)), which first provided global reanalyses using only surface observations. The Twentieth Century Reanalysis project itself is an attempt to provide an observational reanalysis dataset spanning the twentieth century (and prior) to validate climate model output and provide a large sample from which to calculate of daily climate statistics. The model schemes and parameterizations used in the initial release of 20CRv2 are discussed in greater detail in Compo et al. (2011).

The 20CRv2c reanalysis dataset spans years 1851-2014 at six-hourly as well as daily and monthly-averaged intervals. The global grid has a two-degree spatial resolution, providing 91 latitude and 180 longitude grid points at 24 pressure levels ranging from 1000 to 10 hPa. The data is freely available for download at the NOAA-CIRES Twentieth Century Reanalysis Version 2c project webpage, located at https://www.esrl.noaa.gov/psd/data/gridded/data.20thC_ReanV2c.html.

As outlined in Slivinski et al. (2019), the 20CRv2c project produces gridded reanalysis fields by first iterating a coupled weather model against sea surface temperature and sea ice concentration boundary conditions described in Giese et al. (2016) and Hirahara et al. (2014). Next, an Ensemble Kalman Filter technique assimilates pressure observations from the International Surface Pressure Databank (see Cram et al. (2015)) into the model derived fields, producing an adjusted analysis field. As such, four-dimensional fields of atmospheric variables such as air temperature, geopotential

height, vertical velocity, relative humidity, specific humidity and both meridional and zonal wind are produced, all by assimilating only surface pressure values into the model output. More information on the configuration and parameterization schemes used in producing the 20CRv2c reanalysis dataset is found in Appendix A of Slivinski et al. (2019) and at the Twentieth Century Reanalysis Version 2c webpage link above.

While the length of and variables provided by the 20CRv2c dataset provides an extensive basis from which to undertake a statistical study of daily climate data, it is not without problems. In particular, Slivinski et al. (2019) points out two major issues. First, sea-level pressure anomalies before 1870 are too low due to biased ship observations making their way into the surface pressure data assimilation. In addition, a “spectral ringing” artifact was introduced into the precipitation rate field as well as the meridional and zonal wind fields due to an error in downscaling the resolution of the orography used to that of the native model resolution. Despite these limitations, the extensive length and amount of observations contained within the 20CRv2c dataset provide an attractive platform from which to study non-Gaussian climate statistics.

Several atmospheric variables are available for study in the 20CRv2c dataset. From the variables provided, the daily-averaged fields of air temperature (T), geopotential height (Φ'), zonal and meridional wind (u, v) and vertical velocity (ω) were chosen for analysis. Additionally, the relative vorticity (ζ) may be derived from the available variables and will also be investigated. Specifically for this study, the following variables and vertical levels will be analyzed:

- 500 hPa geopotential height
- 950 hPa air temperature
- 300 hPa relative vorticity
- 950 hPa zonal and meridional winds
- 500 hPa vertical velocity

Before the statistical analysis of the 20CRv2c dataset can take place, the time series within the three-dimensional atmospheric fields must first be standardized. In this work, daily averaged 20CRv2c data is used as it omits diurnal cycles. The mean value for each day of the calendar year at each grid point is then calculated. This yields the mean annual cycle, which is subtracted from the daily climate data to produce anomalies for each day of each year at each grid location. The anomalies are then divided by the standard deviation of each location, which yields a time series of standardized anomalies at each grid location. Further processing may be done when seasonal

(DJF, MAM, JJA, SON) or geographical selection of data is desired. This study will look primarily at the persistent wintertime (DJF) statistics of the 20CRv2c dataset, chosen to exclude seasonal cycles as well as to capture the most pronounced synoptic baroclinicity. For the initial evaluation of the SGS distribution, 67 years (6137 days) of persistent wintertime (DJF) 20CRv2c data from 1947-2014 will be analyzed. These results will be used in a comparison to 67 years DJF data from 1880-1947 (6136 days).

The 20CRv2c variables used in this study are now discussed. A more in-depth climatology of the skewness and kurtosis of several atmospheric variables, including the variables analyzed in this work, have been undertaken previously by Sura and Perron (2010), Petoukhov et al. (2008) and most extensively by Perron and Sura (2013).

2.2 500 hPa geopotential height

500 hPa geopotential height is the atmospheric variable that has been studied the most with respect to non-Gaussian statistics. The 500 hPa geopotential height level is by many definitions the level of tropospheric non-divergence, which has implications for forecasting synoptic weather systems and cyclogenesis. The skewness and kurtosis of the 500 hPa geopotential height anomalies was previously discussed in Chapter 1 as an example of large-scale non-Gaussianity, where Figure 1.1 shows the skewness and kurtosis of the 500 hPa DJF 1851-2014 geopotential height anomalies from the 20CRv2c dataset.

2.3 950 hPa air temperature

The 950 hPa air temperature is a level of near-surface solar insolation that warms the land and ocean surfaces. This insolation is generally responsible for the global atmospheric circulation, which when paired with the ocean heat transport comprises the large-scale equator-to-pole energy transport (Perron and Sura (2013)). The skewness and kurtosis of the 950 hPa air temperature anomalies from 1851-2014 are given in Figure 2.1.

The skewness pattern for the near-surface 950 hPa DJF temperature anomalies show a land-ocean relationship, where negatively skewed or near-zero values are found in continental regions and positively skewed values are more concentrated in the ocean basins. Notably, leptokurtic¹ areas

¹Comparisons of excess kurtosis are typically made to the Gaussian distribution, which has a value of zero excess kurtosis. A distribution with zero excess kurtosis may also be called *mesokurtic*. Any distribution with positive

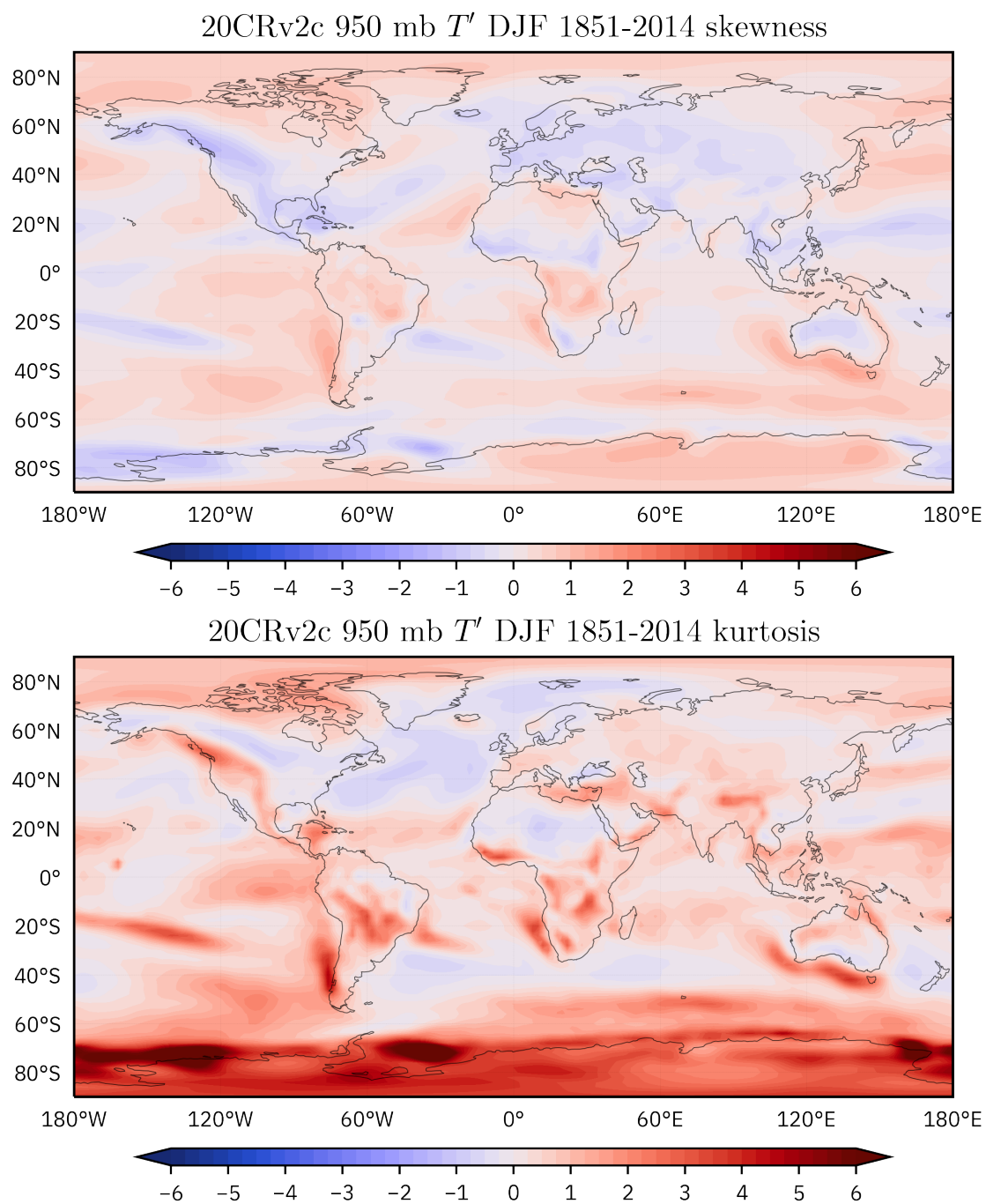


Figure 2.1: Maps of skewness and kurtosis of standardized DJF 950 hPa air temperature anomalies from the 20CRv2c dataset, 1851-2014.

are collocated with large magnitudes of both positive and negative skewness. This relationship is indicative of high and low temperature extremes influencing the statistics of the pdf. Previously, Ruff and Neelin (2012) have shown that non-Gaussian pdf tails exist in station observations of air temperature data while Swanson and Pierrehumbert (1997) found significant skewness in the temperature and meridional velocity fields (and by extension the heat flux) located in the lower troposphere of the Pacific storm track.

2.4 300 hPa relative vorticity

The vorticity is a measure of the rotation in a fluid. The vertical component of the relative vorticity, or $\zeta = \mathbf{k} \cdot (\nabla \times \mathbf{U})$, is associated with storm tracks and locations of cyclonic and anticyclonic flows. In the northern (southern) hemisphere, regions of positive (negative) relative vorticity are coupled with cyclonic weather systems and regions of negative (positive) relative vorticity are associated with anticyclonic weather systems. This makes relative vorticity a useful diagnostic variable (Holton (2004)). Here, the 300 hPa relative vorticity anomalies are investigated as it is the general location of the jet stream. Figure 2.2 gives the skewness and kurtosis of the 20CRv2c 300 hPa 1851-2014 DJF relative vorticity anomalies.

The skewness of the 300 hPa relative vorticity anomalies show a banded latitudinal structure, with sign changes at the equator and at the location of the mean storm track. The kurtosis of the relative vorticity is generally strongly positive, with the exception of bands of negative kurtosis that correspond to the both the mean northern (southern) hemisphere storm track location.

2.5 950 hPa zonal and meridional wind

The near-surface 950 hPa horizontal (zonal and meridional) winds are also investigated in this study. The near-surface wind is an important component to the atmosphere-surface energy exchange e.g., wind-driven ocean currents.

Global skewness and kurtosis patterns for the 950 hPa DJF zonal wind anomalies do not display the same latitudinal banding structure as that of the geopotential height anomalies, save for a single band spanning the Southern Ocean. The skewness map appears to show some features attributed to

excess kurtosis is *leptokurtic*, meaning that the tails of the distribution are fatter. Distributions with thinner tails are termed *platykurtic*, where the excess kurtosis is negative.

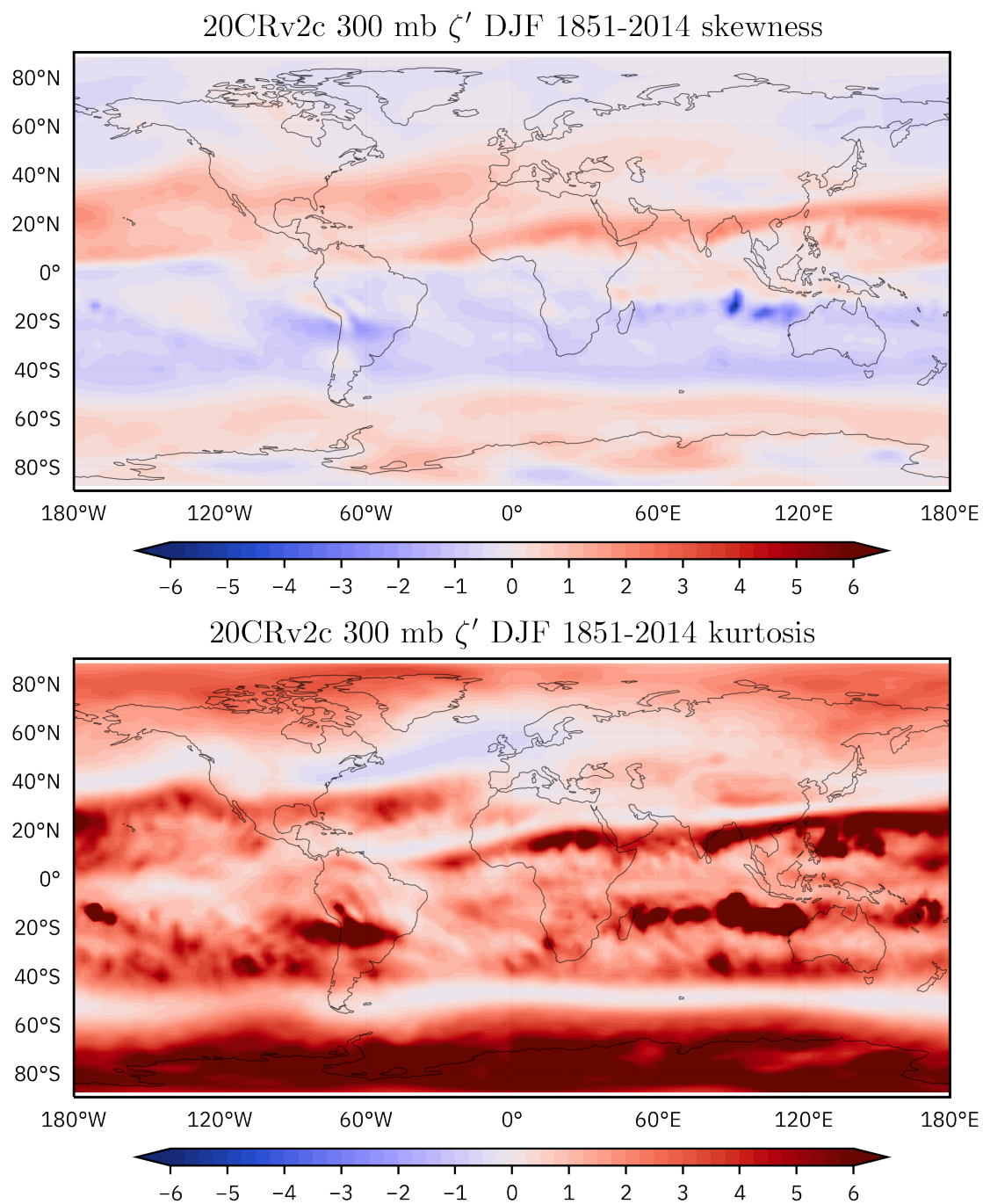


Figure 2.2: Maps of skewness and kurtosis of standardized DJF 300 hPa relative vorticity anomalies from the 20CRv2c dataset, 1851-2014.

synoptic scale weather patterns, but the kurtosis map is much less structured. As the map in Figure 2.3 is of the near-surface zonal wind, it makes sense that natural geographical features have an effect on the spatial distribution of the higher moments. For example, the topography of the Himalayan mountain system appears to perturb both the skewness and kurtosis fields. Positive skewness values near the equator are in a region dominated by the Hadley cell, a major tropospheric feature of the general atmospheric circulation, where the zonal wind is typically easterly. Finally, the magnitude of the skewness appears larger over the ocean than land.

Given by Figure 2.4, the skewness of the 950 hPa meridional wind is highly dependent on the hemisphere. In boreal winter (DJF), the northern hemisphere typically shows positively skewed values, while the southern hemisphere indicates generally negative values. This corresponds with the general equator-to-pole circulation. The kurtosis is generally positive near the equator, but becomes less positive towards the poles save for pockets of positive kurtosis over Antarctica. Perron and Sura (2013) suggests that such a pattern of 950 hPa meridional wind skewness and kurtosis may be indicative of a non-Gaussian poleward energy transport.

2.6 500 hPa vertical velocity

The vertical component of the wind velocity is investigated here in pressure coordinates. The vertical component of the wind describes the upward and downward motion of the wind, which is related to convection, instability and orographic lift. The vertical velocity at 500 hPa is chosen as it is the level of non-divergence, where the vertical velocity is typically at a maximum (Perron and Sura (2013)).

In Figure 2.5, a much different picture of non-Gaussianity emerges compared to the prior variables. The near-zero or negative skewness located over most of the globe is seemingly indicative of a general instability of the atmosphere. Few regions outside of Antarctica and the polar northern hemisphere feature positive kurtosis. Meanwhile, the kurtosis of the vertical velocity is positive everywhere, with more neutral values present in regions of storm tracks and a small equatorial band. Perron and Sura (2013) explains the skewness and kurtosis patterns of the vertical velocity by suggesting that vertical motion is typically restricted to either a “common stable mode” with stable, slowly descending air and a “rarer unstable mode” where air trapped in a convective burst is propelled up in the atmosphere at faster speed.

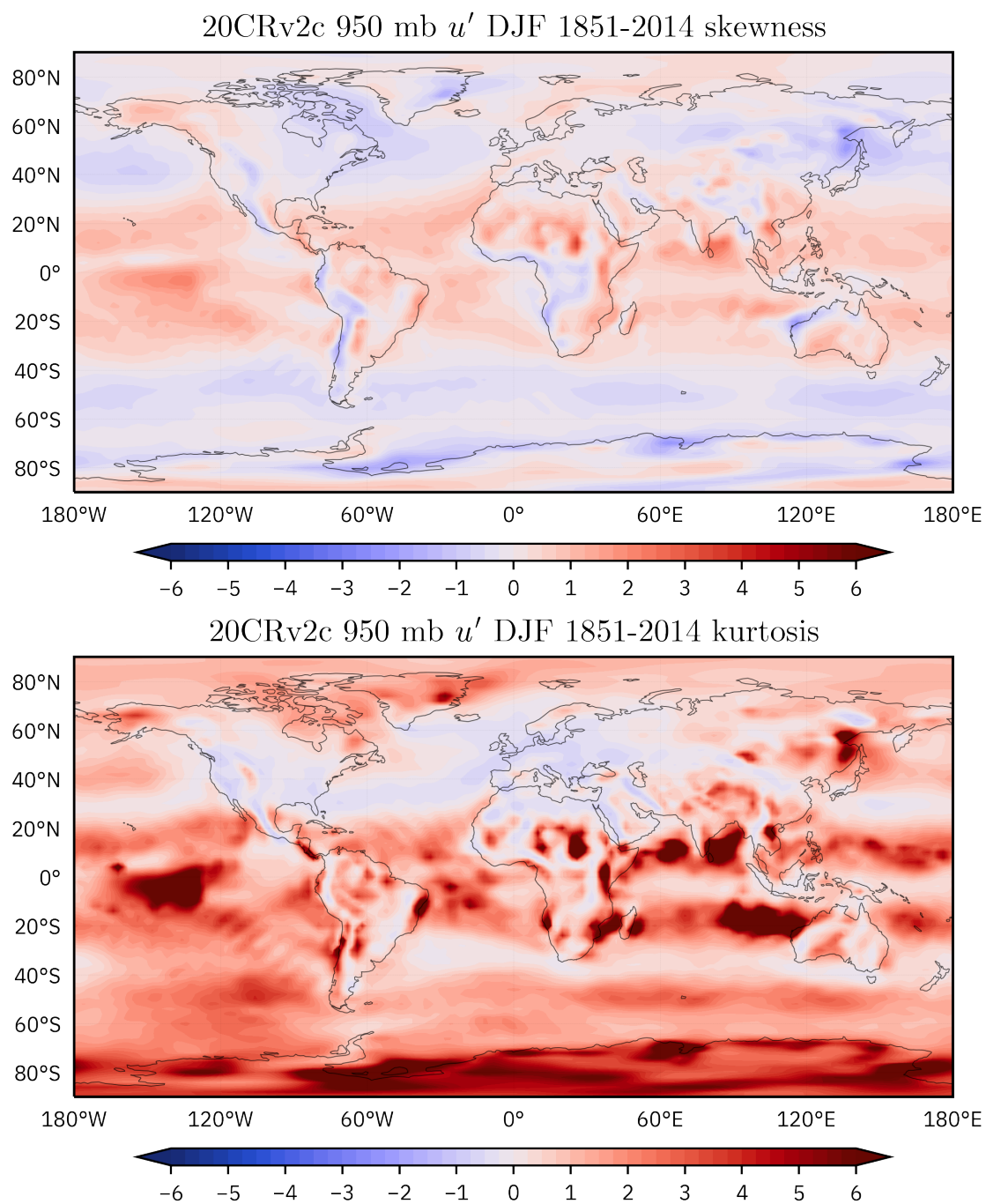


Figure 2.3: Maps of skewness and kurtosis of standardized DJF 950 hPa zonal wind anomalies from the 20CRv2c dataset, 1851-2014.

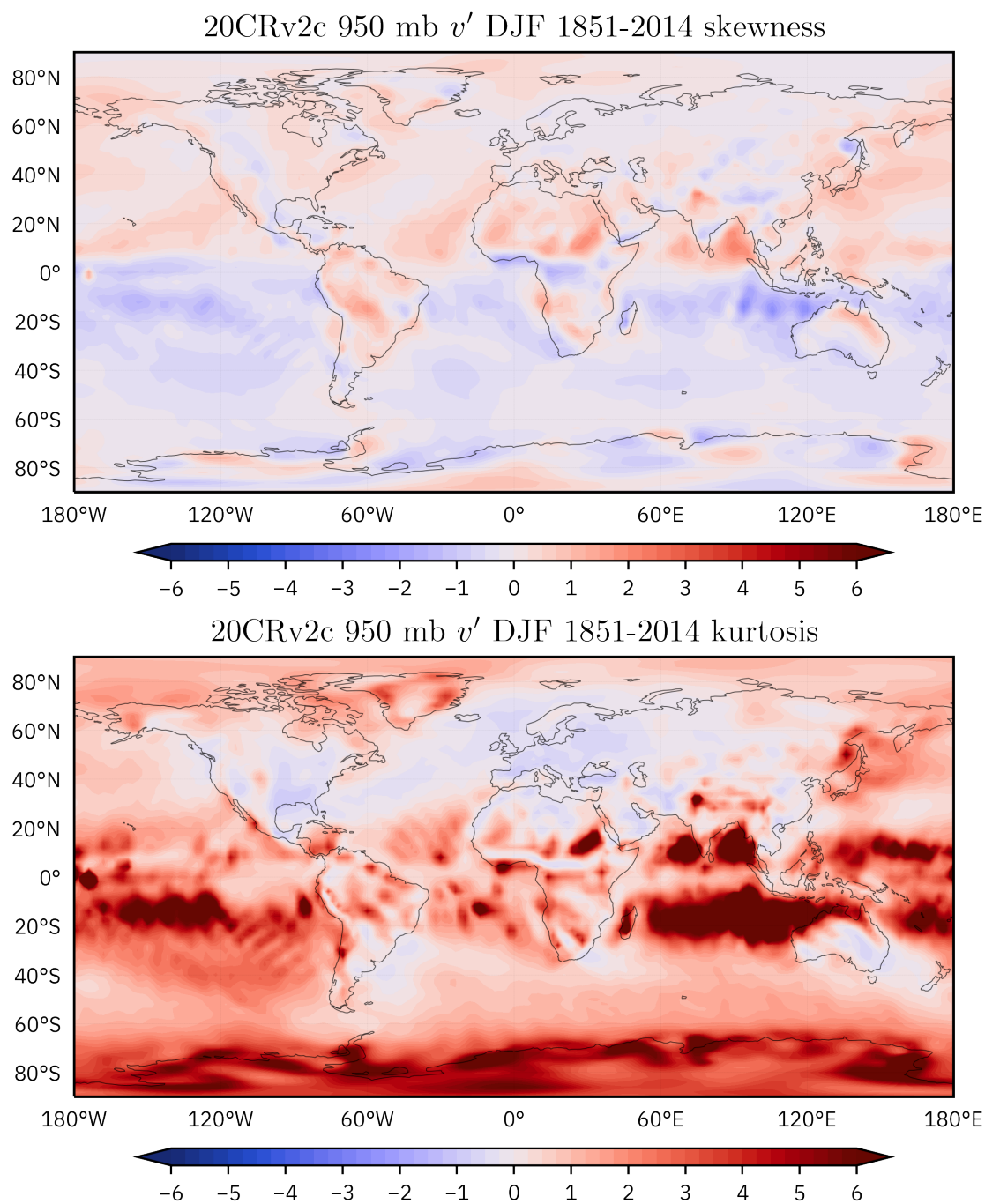


Figure 2.4: Maps of skewness and kurtosis of standardized DJF 950 hPa meridional wind anomalies from the 20CRv2c dataset, 1851-2014.

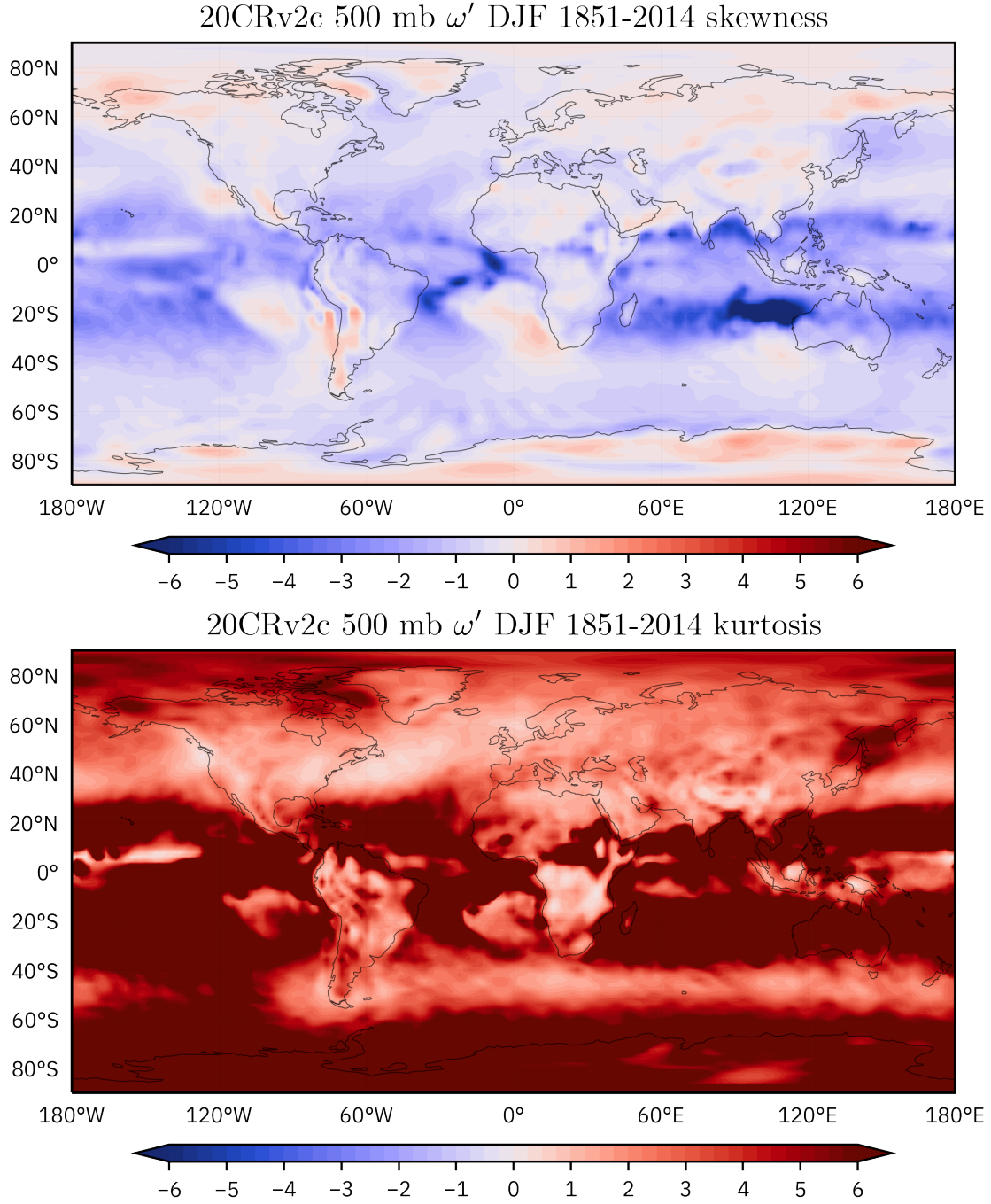


Figure 2.5: Maps of skewness and kurtosis of standardized DJF 500 hPa vertical velocity anomalies from the 20CRv2c dataset, 1851-2014.

CHAPTER 3

SOME STOCHASTIC CLIMATE THEORY

Motivated by Hasselmann’s climate model, the correlated additive and multiplicative (CAM) noise stochastic climate model produces both the red climate spectrum and non-Gaussian statistics seen in daily atmospheric data sets.

3.1 Physical motivation

Stochastic modeling of the climate system is based on the assumption that distinct physical processes are represented by distinct spatiotemporal scales. However, the intrinsic non-linearity of the climate system is due to the spatiotemporal interactions of each scale with other scales – any strict separation of time scales is physically impossible. Franzke et al. (2015) highlights the wave tank experiment of Williams et al. (2003), which shows that a two-layer stratified fluid in a rotating annulus can experience Rossby wave regime transitions by the influence of inertia-gravity waves. Subsequent numerical modeling by Williams et al. (2004) showed that the large scale wave regime transitions in an quasi-geostrophic numerical model were only achieved when the inertia-gravity waves were included as a stochastic forcing.

Franzke et al. (2015) points out that for many applications, predictable phenomena of practical interest are confined to a few larger scales¹, while the small scale processes (including but not limited to turbulence) often remain unpredictable. This seems an appropriate characterization whether considering the problem of resolving turbulence in numerical models or the evolution of inertial-gravity waves in a rotating annulus. As such, representation of small scale processes in models typically involves some kind of parameterization. Ultimately, the time scale separation employed in stochastic climate models is a useful approximation to the non-linear interactions of the true climate system.

¹Generally, the larger and slower scales, such as the climate or regime behavior, are more easily resolved and evolved in models. Franzke et al. (2015) points out that numerical models spend a majority of their computational time resolving the faster, smaller scales.

Before moving to discuss particular stochastic climate models and their statistics, some mathematical preliminaries of the stochastic dynamics are discussed.

3.2 The stochastic differential equation

The stochastic dynamics of any scalar value, x , may be described by the stochastic differential equation (SDE)

$$\frac{dx}{dt} = A(x) + B(x) \eta(t) , \quad (3.1)$$

where $A(x)$ represents a slow, deterministic process and the product of $B(x)$ and $\eta(t)$ is the stochastic approximation to a fast phenomenon². $\eta(t)$ is defined as the delta-correlated white noise for all time, t . For the independent, Gaussian process $\eta(t)$,

$$\begin{aligned} \langle \eta(t) \rangle &= 0 \\ \langle \eta(t) \eta(t') \rangle &= \delta(t - t') , \end{aligned}$$

where the time average is denoted as $\langle \dots \rangle$ and δ is the Dirac delta-function. As $\eta(t)$ is not a continuous function, Equation (3.1) cannot be differentiated. To evaluate Equation (3.1) continuously, one must first integrate.

$$x(t) - x(t_0) = \int_{t_0}^t A[x(t')] dt' + \int_{t_0}^t B[x(t')] \eta(t') dt' . \quad (3.2)$$

However, $\eta(t)$ is still a discontinuous function, even though the white noise is defined at each time step. To help with the integration, $\eta(t)$ can be replaced by a continuous integral by invoking the Wiener process: the time derivative of the white noise (Evans (2013)). Doing so gives

$$x(t) - x(t_0) = \int_{t_0}^t A[x(t')] dt' + \int_{t_0}^t B[x(t')] dW(t') , \quad (3.3)$$

where the Wiener process is defined as

$$W(t) = \int_0^t \eta(t') dt' , \quad (3.4)$$

or, defined only in the integral itself

$$dW(t) = \eta(t) dt . \quad (3.5)$$

²See Sornette (2006), Sura and Sardeshmukh (2008), Sardeshmukh and Sura (2009), Sura (2011) for a similar development.

To distinguish between stochastic and deterministic calculus, the scaling of the time increment, dt , is compared with the Wiener process. By substituting Equation (3.4) and simplifying, the incrementation of the Wiener process may be written as

$$\begin{aligned} dW(t) &= W(t + dt) - W(t) \\ &= \int_t^{t+dt} \eta(t') dt' . \end{aligned} \quad (3.6)$$

Here, Sura (2011) advocates that using the definition of $dW(t)$ can be used to find the auto-covariance of two time values, t and s , over a time step, dt . Using Equation (3.6), the auto-covariance of the Wiener process, given as $\langle dW(t)dW(s) \rangle$, may be expressed as

$$\begin{aligned} \langle dW(t)dW(s) \rangle &= \left\langle \int_t^{t+dt} \int_s^{s+dt} \eta(t') \eta(s') dt' ds' \right\rangle \\ &= \int_t^{t+dt} \int_s^{s+dt} \langle \eta(t') \eta(s') \rangle dt' ds' \\ &= \int_t^{t+dt} \int_s^{s+dt} \delta(t' - s') dt' ds' \\ &= dt \delta_{t,s} . \end{aligned} \quad (3.7)$$

$\delta_{t,s}$ is defined here as the Kronecker delta, where $\delta_{t,s} = 0$ when $t \neq s$ and $\delta_{t,s} = 1$ if $t = s$. In the instance that $\delta_{t,s} = 1$, Equation (3.7) simplifies to

$$\langle dW(t)^2 \rangle = dt . \quad (3.8)$$

This expression states that the standard deviation of the Wiener process (the white noise term of the stochastic approximation to fast phenomena) scales with \sqrt{dt} . This scaling is a crucial difference from deterministic calculus, which scales by dt . Furthermore, a calculation by Gardiner (2009) shows that Equation (3.8) is valid even without the ensemble average. This means that

$$dW(t)^2 = dt , \quad (3.9)$$

which is an expression that will be used in subsequent calculations.

As the properties of the stochastic differential equation of a scalar value, x , have been considered, let us now consider the stochastic differential equation of a function of x , such as $f(x)$. Following Gardiner (2009), one may use a Taylor expansion in $dW(t)$ to express the derivative of $f(x)$, assuming that $f(x)$ is twice differentiable (Øksendal (2007)). This yields

$$\begin{aligned} df(x) &= f(x + dx) - f(x) \\ &= f'(x) dx + \frac{1}{2} f''(x) dx^2 + \dots + \frac{1}{n} f^n(x) dx^n . \end{aligned}$$

Taking the stochastic differential equation seen in Equation (3.1) and substituting the definition of the Wiener process used in Equation (3.5) gives

$$dx = A(x) dt + B(x) dW(t) . \quad (3.10)$$

Substitution of this expression into Equation (3.2) for dx creates

$$df(x) = f'(x) [A(x) dt + B(x) dW(t)] + \frac{1}{2} f''(x) B(x)^2 dW(t)^2 , \quad (3.11)$$

where $dW(t)^2 = dt$ and the terms of order two or greater are neglected. As the scaling of the Wiener process is equivalent to the scaling of the deterministic calculus (see Equation (3.8)), Equation ((3.11) may be rearranged to obtain Itô's formula, written as

$$df(x) = \left[f'(x) A(x) + \frac{1}{2} f''(x) B(x)^2 \right] dt + f'(x) B(x) dW(t) , \quad (3.12)$$

which illustrates that changing variables in stochastic calculus is not the same as doing so in deterministic calculus (Gardiner (2009)).

3.3 The Fokker-Planck equation

After analyzing the stochastic differential equation of an arbitrary function, $f(x)$, and achieving Itô's formula, the ultimate goal is to obtain an expression that is able to describe the evolution of the function $f(x)$ with time. This expression is called the Fokker-Planck equation, which describes the time evolution of the probability density function of a stochastic variable under deterministic (non-random) and diffusive effects (Risken (1996), Øksendal (2007), Gardiner (2009)).

To invoke the Fokker-Planck equation, consider the evolution in time of an arbitrary function, denoted as $\frac{\langle df(x) \rangle}{dt}$, where $\langle \dots \rangle$ represents the ensemble average. Gardiner (2009) suggests substituting Itô's formula in place of the time derivative of $f(x)$ so that the time evolution of a stochastic function may be represented as

$$\begin{aligned} \frac{\langle df(x) \rangle}{dt} &= \frac{d}{dt} \langle f(x) \rangle \\ &= \left\langle \frac{\partial f(x)}{\partial x} A(x) + \frac{1}{2} \frac{\partial^2 f(x)}{\partial x^2} B(x)^2 \right\rangle . \end{aligned} \quad (3.13)$$

Upon the realization that the scalar value x has a probability density function, $p(x)$, Equation (3.13) may be re-written as

$$\frac{d}{dt} \langle f(x) \rangle = \int_{-\infty}^{\infty} \frac{\partial p(x)}{\partial t} f(x) dx , \quad (3.14)$$

since the expected value (mean) of an arbitrary variable may be defined as

$$\langle x \rangle = \int_{-\infty}^{\infty} x p(x) dx . \quad (3.15)$$

Substituting in the value of $f(x)$ from Equation (3.13) into Equation (3.14) results in

$$\int_{-\infty}^{\infty} \frac{\partial p(x)}{\partial t} f(x) dx = \int_{-\infty}^{\infty} \left[\frac{\partial f(x)}{\partial x} A(x) + \frac{1}{2} \frac{\partial^2 f(x)}{\partial x^2} B(x)^2 \right] p(x) dx . \quad (3.16)$$

After integrating by parts, omitting surface terms and removing the arbitrary function $f(x)$ from both sides of the equation (see Gardiner (2009) for more details), the one-dimensional Itô Fokker-Planck equation is achieved:

$$\frac{\partial p(x, t)}{\partial t} = -\frac{\partial}{\partial x} A(x) p(x, t) + \frac{1}{2} \frac{\partial^2}{\partial x^2} B(x)^2 p(x, t) . \quad (3.17)$$

More generally, Sardeshmukh and Sura (2009) write that the n -dimensional Fokker-Planck equation of a stochastic variable, $\vec{x}(t)$, is

$$\begin{aligned} \frac{\partial p(\vec{x}, t)}{\partial t} = & - \sum_i \frac{\partial}{\partial x_i} \left[A_i + \hat{\alpha} \sum_{j,k} \left(\frac{\partial}{\partial x_j} B_{ik} \right) B_{jk} \right] p(\vec{x}, t) \\ & + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} (\mathbf{B}\mathbf{B}^T)_{ij} p(\vec{x}, t) , \end{aligned} \quad (3.18)$$

where $p(\vec{x}, t)$ is the probability density function.

Gardiner (2009) writes that the phenomena described by the Fokker-Planck equation may be thought of as a kind of “diffusion process,” as the equation itself is composed of drift and diffusion terms. In the general sense, A_i represents the deterministic term, also referred to as the deterministic drift, while the second term in the square brackets is designated as the noise-induced drift. The last term in Equation (3.18) is the contribution of noise to the diffusion of the probability density (Sura (2011)). However, the most important distinction to be made in the Fokker-Planck equation is that the value of $\hat{\alpha}$ determines the appropriate application of the Fokker-Planck equation to a given physical system. Though there are different values that $\hat{\alpha}$ may take, two are valued for their ability to approximate different kinds of physical systems³. These two $\hat{\alpha}$ values also correspond to different types of stochastic calculus that are suited for use in either discrete or continuous cases.

³Øksendal (2007) provides a thorough discussion of the two “useful” types of stochastic calculus.

In the case that $\hat{\alpha} = \frac{1}{2}$, the calculus is Stratonovich, while if $\hat{\alpha} = 0$, Itô calculus is used and the noise-induced drift is zero. Stratonovich calculus is suited for use in continuous systems (such as the atmosphere-ocean) where variables with small time correlations may be approximated as Gaussian white noise. Systems that are approximated from discrete data will instead benefit from Itô calculus, where the total drift is the combination of the deterministic and the noise-diffusive drift terms (Sura (2011)). For example, Itô calculus is used in Equation (3.17), where $A(x)$ is defined as the combination of the deterministic and noise-diffusive drift terms, also known as the effective drift.

3.4 Stochastic dynamics of climate variability

After briefly reviewing some principles of stochastic calculus and dynamics, the task is now to assemble equations that describe the stochastic approach to climate variability. Here, an assumption is made that climate dynamics may be divided into two different time scales. As such, the slow, deterministic process (climate) and the stochastic approximation of the fast phenomena (weather) may be described with a stochastic differential equation. The fast term is approximated using Gaussian white noise, used to ensure that the fast process decorrelates more rapidly than the slow process. The mathematical framework of the time scale decomposition was previously formulated in Equation (3.1), written again following Sura (2011) as

$$\frac{d\vec{x}}{dt} = \vec{A}(\vec{x}, t) + \mathbf{B}(\vec{x}, t) \vec{\eta}(t) , \quad (3.19)$$

where vector $\vec{A}(\vec{x})$ represents a slow, deterministic process and the product of the matrix $\mathbf{B}(\vec{x})$ and $\vec{\eta}(t)$ is the stochastic approximation to a fast phenomenon. In the following equations, the stochastic approximation of the fast phenomena will be represented as $\mathcal{B}(\vec{x})$, per the notation used by Sura (2011) and the definition $\mathcal{B}(\vec{x}) = \mathbf{B}(\vec{x}) \vec{\eta}(t)$ where the time dependencies of \vec{A} and \mathcal{B} will be implied. As such, performing a Reynolds decomposition of \vec{x} into its mean, $\langle \vec{x} \rangle$, and anomaly, \vec{x}' , permits a Taylor's expansion in linearizing Equation (3.19) about the mean. Doing so yields

$$\frac{d(\langle \vec{x} \rangle + \vec{x}')}{dt} = \vec{A}(\langle \vec{x} \rangle) + \left. \frac{\partial \vec{A}(\langle \vec{x} \rangle)}{\partial \langle \vec{x} \rangle} \right|_{\vec{x}=\langle \vec{x} \rangle} (\vec{x} - \langle \vec{x} \rangle) + \mathcal{B}(\langle \vec{x} \rangle) + \left. \frac{\partial \mathcal{B}(\langle \vec{x} \rangle)}{\partial \langle \vec{x} \rangle} \right|_{\vec{x}=\langle \vec{x} \rangle} (\vec{x} - \langle \vec{x} \rangle) + \mathcal{R} \quad (3.20)$$

where a residual, \mathcal{R} , is included to account for omitted terms, specifically the non-linear terms of the Taylor expansion. In Equation (3.20), the respective first-order partial derivatives of \vec{A} and \mathcal{B}

are Jacobian matrices, which will be written as $\mathbf{J}_{\vec{A}}$ and $\mathbf{J}_{\mathcal{B}}$. Simplifying Equation (3.20) yields the time tendency of the anomaly, \vec{x}' , written as

$$\frac{d(\vec{x}')}{dt} = \mathbf{J}_{\vec{A}}(\langle \vec{x} \rangle)(\vec{x}') + \mathcal{B}(\langle \vec{x} \rangle) + \mathbf{J}_{\mathcal{B}}(\langle \vec{x} \rangle)(\vec{x}') + \mathcal{R} . \quad (3.21)$$

In this instance, it is assumed that the time tendency and the deterministic process effectively cancel at the mean. Moreover, the partial derivative of the multiplicative noise term with respect to the mean, $\mathbf{J}_{\mathcal{B}}(\langle \vec{x} \rangle)$, is not guaranteed to be zero (Sura (2011)). This is due to the linearization, where the non-linear higher-order terms were neglected. Therefore, an additional term representing a corrective mean forcing is needed to ensure that the mean of the anomaly is zero, which results in

$$\frac{d(\vec{x}')}{dt} = \mathbf{J}_{\vec{A}}(\langle \vec{x} \rangle)(\vec{x}') + \mathcal{B}(\langle \vec{x} \rangle) + \mathbf{J}_{\mathcal{B}}(\langle \vec{x} \rangle)(\vec{x}') - \langle \mathbf{J}_{\mathcal{B}}(\langle \vec{x} \rangle)(\vec{x}') \rangle + \mathcal{R} . \quad (3.22)$$

Finally, Equation (3.22) is rewritten into the notation used by Sura (2011), which makes use of Einstein's summation convention while combining terms into tensors and omitting the prime notation. This type of notation is used in the quantitative stochastic models put forth by Hasselmann (1976) and Sardeshmukh and Sura (2009), both of which are described in the following sections. This rewriting yields

$$\frac{dx_i}{dt} = A_{ij}x_j + (G_{ip} + E_{ijp}x_j)\eta_p - D_i + r_i . \quad (3.23)$$

The first term on the right hand side of Equation (3.23) is the representation of the linear, deterministic Jacobian matrix of Equation (3.22), while $G_{ip}\eta_p$ and $E_{ijp}x_j\eta_p$ denote the additive noise and multiplicative noise terms, respectively. D_i is the corrective mean noise forcing vector inserted into Equation (3.22), shown by Sardeshmukh and Sura (2009) to be equal to $-\frac{1}{2}E_{ijp}G_{jp}$. As written, Equation (3.23) is a foundation for the application of stochastic climate models. Note that from this equation, both Gaussian and non-Gaussian statistics can be represented. In the case that the multiplicative noise is equal to zero ($E_{ijp} = 0$), only the additive noise terms survive. Such a model would only be able to capture the Gaussian statistics of a climate system (Gardiner (2009), Sura (2011)). This type of approach is seen in Hasselmann's quantitative model. Conversely, the system where the multiplicative noise is not zero ($E_{ijp} \neq 0$) will feature non-Gaussian statistics (Sardeshmukh and Sura (2009), Sura (2011)). As such, Hasselmann's quantitative model will be examined first before describing the multiplicative noise system in detail. Though Hasselmann's model is only able to produce Gaussian statistics, it is nonetheless the foundation for achieving and understanding the multiplicative noise model.

3.5 Hasselmann's climate model

The major threads of current stochastic climate research may be traced back to Hasselmann (1976), where the dynamics of the climate system are expressed as contributions from both a rapidly decorrelating process and a slowly decorrelating process. Consider the reduction of Equation (3.23) to that of a one-dimensional expression where the multiplicative noise term, E_{ijp} , is equal to 0, only the additive noise term persists. Further, after removing the tensor notation and residuals from Equation (3.23), the result is Hasselmann's quantitative model for climate variability

$$\frac{dx}{dt} = -\lambda x + \sigma \eta, \quad (3.24)$$

where, using the notation of Sura (2011), a positive damping coefficient, λ , has been substituted for A_{ij} and the constant amplitude of the Gaussian white noise, σ , is exchanged for G_{ip} . η is the Gaussian white noise that represents rapidly varying weather fluctuations. The time average of the Gaussian white noise is denoted by $\langle \dots \rangle$ and is equal to zero while $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$ for time t and lag t' . This first-order autoregression model has been used to describe characteristics of SST variability⁴ and it is also the familiar expression that describes Brownian motion⁵.

There are two particularly interesting results that may be obtained from Hasselmann's climate model. First, by calculating the autocorrelation function $\rho(t')$ for lag t' , one achieves

$$\rho(t') = \frac{\langle x(t)x(t+t') \rangle}{\sigma^2} = e^{-\lambda t'}, \quad (3.25)$$

which reveals a red-noise (Brownian-noise) frequency spectrum, where lower frequencies have increasingly larger intensities. In the climate system, this red-noise spectrum is analogous to the damping of high frequency events by the low frequency components, which in turn creates the slowly decorrelating response often seen in climate data (Sura (2011)).

Second, the stationary probability density function of Equation (3.24), $p(x)$, is found by invoking the Fokker-Planck equation, written as

$$\frac{d}{dx} \lambda x p(x) + \frac{1}{2} \frac{d^2}{dx^2} \sigma^2 p(x) = 0. \quad (3.26)$$

By integrating away one derivative and assuming a solution of the form

$$p(x) = N \exp(f(x))$$

⁴See Reynolds (1978) for an example.

⁵Gardiner (2009) contains an interesting history on the development of Brownian motion.

where N is a constant, the solution of the second order ordinary differential equation is

$$p(x) = \frac{1}{\sqrt{2\pi(\sigma^2/2\lambda)}} \exp\left[-\frac{x^2}{2(\sigma^2/2\lambda)}\right], \quad (3.27)$$

i.e., a probability density function of a Gaussian distribution, centered at zero with variance $\langle x^2 \rangle = \sigma^2/2\lambda$ (Sura (2011)). This Gaussian pdf, however, is at odds with the statistics observed in non-Gaussian climate data (see Figure 1.1, Chapter 2 and Perron and Sura (2013)). Therefore, while Hasselmann's model is able to produce the now familiar red-noise climate spectrum, it is unable to reconstruct the non-Gaussian statistics that are seen in climate variability. In addition, Sura (2011) demonstrates that modifying the additive stochastic forcing to be non-Gaussian is not enough to induce a non-Gaussian pdf response. To account for the non-Gaussian statistics that are seen in observations, one must introduce multiplicative noise into the stochastic model.

3.6 Correlated additive and multiplicative noise model

To reproduce the non-Gaussian climate statistics observed in daily climate observations, Sura and Sardeshmukh (2008) and Sardeshmukh and Sura (2009) supplemented the stochastic climate model of Hasselmann (1976) with multiplicative noise. This process has been outlined previously in Section 3.4⁶, culminating in the formulation of Equation (3.23).

3.6.1 The CAM noise pdf

Consider again the one-dimensional case where the residual terms are neglected, (3.23) can be expressed as a Stratonovich SDE⁷ given by

$$\frac{dx}{dt} = Ax + b\eta_1 + (Ex + g)\eta_2 - \frac{1}{2}Eg. \quad (3.28)$$

Here, the Stratonovich stochastic forcing is contained within the additive noise forcing terms, $b\eta_1$ and $g\eta_2$, and the multiplicative noise term, $Ex\eta_2$, where Sardeshmukh et al. (2015) remark that the $(Ex + g)\eta_2$ is the CAM noise described in Sardeshmukh and Sura (2009). As before, the Gaussian white-noise terms η_1 and η_2 are δ -correlated with a time average of zero.

⁶The derivation of the CAM noise model is described at length in the works of Sura and Sardeshmukh (2008), Sardeshmukh and Sura (2009) and provided additional context in Sura (2011).

⁷Here the notation of Sardeshmukh and Sura (2009) and Sardeshmukh et al. (2015) is again used.

Analogous to Hasselmann's quantitative model, the multiplicative noise non-Gaussian SDE given by Equation (3.28) also yields a red frequency spectrum and the stationary probability density function, $p(x)$, may again be achieved by invoking the Fokker-Planck equation:

$$p(x) = \frac{1}{\mathcal{N}} \left[(Ex + g)^2 + b^2 \right]^{-(1+(1/E^2))} \exp \left[\frac{2g}{E^2 b} \arctan \left(\frac{Ex + g}{b} \right) \right]. \quad (3.29)$$

This pdf is designated the stochastically generated skewed (SGS) distribution. The notation in Equation (3.29) is of Sardeshmukh et al. (2015), where the normalization constant, \mathcal{N} , is provided such that the pdf will integrate to one and is defined as

$$\mathcal{N} = \frac{2\pi\nu^{1/2} (2b)^{-(2\nu+1)} \Gamma(2\nu+1)}{\Gamma(\nu+1-iq/2) \Gamma(\nu+1+iq/2)}. \quad (3.30)$$

For convenience, $\nu = 1/E^2$ and $q = 2g\nu/b$ and Γ is the special gamma function which accepts a complex argument.

3.6.2 Statistical moments

Statistical moments characterize the location, shape and symmetry of a probability density function. Defined previously in Equation (1.2), the mean, or the first central moment, $\langle x^1 \rangle$, is standardized here to be zero. The second central moment, $\langle x^2 \rangle$, or variance, is a measure of the spread of the probability distribution around the mean. Though variance is not a property of Gaussianity per se, it is useful in providing a picture of the range of data. The standard deviation, σ , is the square root of the variance.

Skewness, or the third moment, $\langle x^3 \rangle$, describes the symmetry of the probability distribution. Negative skewness corresponds to a larger left tail of the probability distribution than a pdf with the same mean and variance, while positive skewness coincides with a larger tail on the right side of the pdf, compared to a pdf with the same mean and variance. The fourth moment, $\langle x^4 \rangle$, is called kurtosis and is a measure of the “peakedness” of the pdf or the amount of the probability distribution located in the tails. Positive kurtosis corresponds to a “taller” peak of the pdf around the mean as well as a larger amount of data in the tails of the pdf. Conversely, negative kurtosis is seen in pdfs that have less data in the tails and a “broader” cluster of the probability distribution located about the mean. Negative and positive kurtosis can also be defined in relation to the Gaussian (mesokurtic) distribution. The “broader” pdf exhibited by negative kurtosis may be called platykurtic, where less data is found in the tails and near the mean. Likewise, the taller pdf obtained

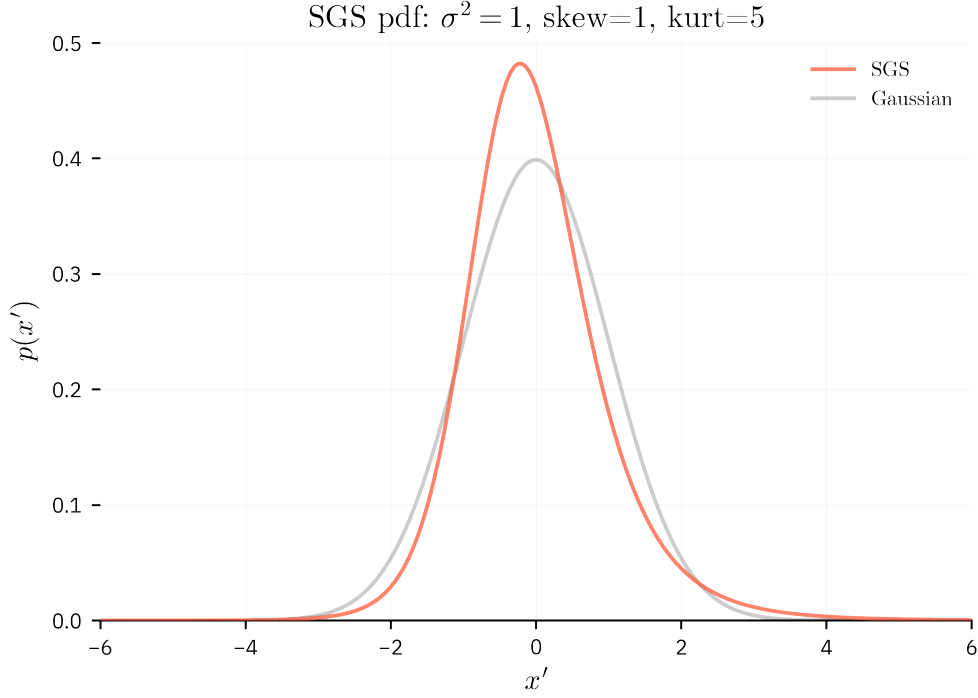


Figure 3.1: Example SGS pdf (red) with variance and skewness of one, kurtosis of five. The standard normal distribution (gray) is provided for reference.

via positive kurtosis may be defined as leptokurtic, where the data is clustered around both the mean and tails. Note that in order to compare values of skewness and kurtosis from different pdfs, the moment's value is normalized as described in Equation (1.3).

An example SGS distribution is shown in Figure 3.1, where the variance and skewness are unity and the kurtosis is five. The SGS pdf in red is contrasted with the standard normal pdf in grey. Here, the positive kurtosis of the SGS distribution is evident from the peakedness and heavy tails. The positive skewness is also shown by the slight negative shift of the pdf peak along with the elongated positive tail.

To produce an expression for the n^{th} moment of the CAM noise model, one first obtains the Fokker-Planck equation associated with the SDE given by Equation (3.28), then multiplies by x^{n-1} and integrates the resulting equation by parts over x (Sura and Sardeshmukh (2008), Sardeshmukh

and Sura (2009)). Doing so yields the first four moments:

$$\begin{aligned}
\langle x \rangle &= \mu = 0 \\
\langle x^2 \rangle &= \sigma^2 = \frac{g^2 + b^2}{2 - E^2} \\
\langle x^3 \rangle &= \text{skew} = \frac{2Eg}{\sigma(1 - E^2)} \\
\langle x^4 \rangle &= \text{kurt} = \frac{3}{2} \left[\frac{1 - E^2}{1 - (3/2)E^2} \right] \text{skew}^2 + \left[\frac{3E^2}{1 - (3/2)E^2} \right].
\end{aligned} \tag{3.31}$$

From these higher moments, Sura and Sardeshmukh (2008) show that a unique link exists between skewness and kurtosis. By solving for kurtosis as a function of skewness, a parabolic relationship is obtained, given by the expression

$$\text{kurt} \geq \left(\frac{3}{2} \right) \text{skew}^2. \tag{3.32}$$

In practice, an empirically estimated vertical offset must generally be applied to the right side of Equation (3.32), giving

$$\text{kurt} \geq \left(\frac{3}{2} \right) \text{skew}^2 - r. \tag{3.33}$$

where r is a non-negative constant. This vertical offset is required to reconcile the small negative bias that exists when the skewness-kurtosis relationship is applied to empirical observations. So far, a general physical explanation for the skewness-kurtosis relationship and the slight negative bias remains elusive.

Figure 3.2 shows the parabolic skewness-kurtosis relationship for daily wintertime (DJF) standardized 950 hPa zonal wind anomalies taken from the 20CRv2c dataset from 1947-2014. This relationship is demonstrated as a two-dimensional histogram of skewness as a function of kurtosis, where the bins represent the sum of the calculated skewness and kurtosis for the time series at each of the 10512 grid points. The skewness-kurtosis constraint given in Equation (3.33) is shown with the black parabola, where $r = 0.972$. The light gray parabolic curve represents the threshold $\text{kurt} = \text{skew}^2 - 2$, below which the kurtosis does not exist. This $\text{kurt} = \text{skew}^2 - 2$ constraint is valid for sample data drawn from any pdf (Wilkins (1944)).

It is apparent from Figure 3.2 that a well-defined bound of non-Gaussianity exists in the zonal wind anomalies taken from the reanalysis data, as nearly all the points are contained within the parabola described by Equation (3.32). Similar parabolic relationships also exist for other atmospheric-oceanic quantities (see Sura and Sardeshmukh (2008), Sardeshmukh and Sura (2009),

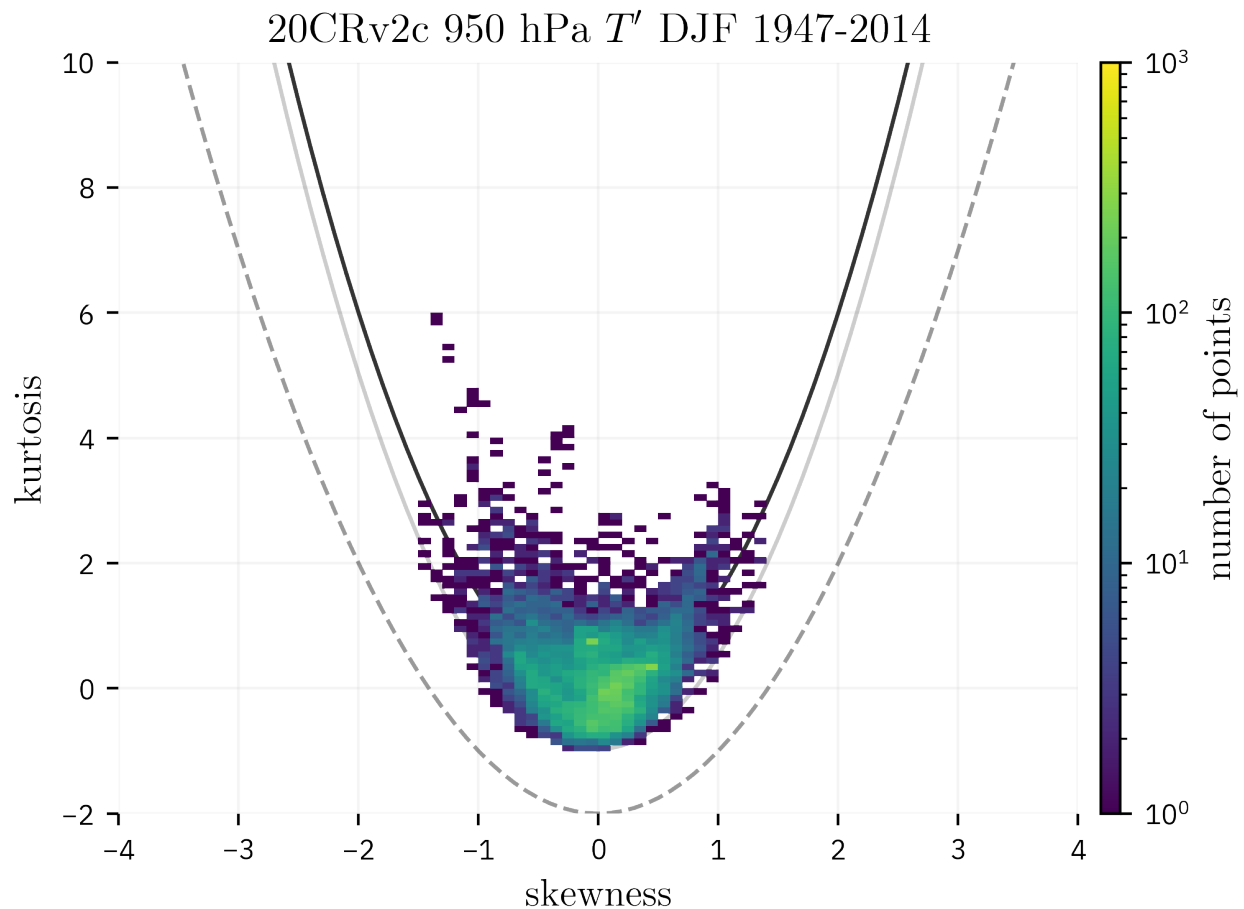


Figure 3.2: Scatterplot of kurtosis versus skewness showing standardized 950 hPa air temperature anomalies (T') for wintertime (DJF) 20CRv2c data (1947-2014). The dark solid parabola indicates the skewness-kurtosis relationship $\text{kurt} \geq \left(\frac{3}{2}\right) \text{skew}^2$, where the light solid parabola is $\text{kurt} \geq \left(\frac{3}{2}\right) \text{skew}^2 - 0.972$. The dotted line represents the less stringent $\text{kurt} = \text{skew}^2 - 2$ constraint.

Sura and Perron (2010) for more examples). Moreover, the addition of multiplicative noise to Hasselmann's stochastic climate model provided by Sura and Sardeshmukh (2008) appears to closely represent the non-Gaussian statistics of atmosphere-ocean phenomena. Note that the vertical offset r in Figure 3.2 has been determined empirically from the data.

Additionally, Sardeshmukh and Sura (2009) demonstrated that a stochastic climate model featuring multiplicative noise will produce power law distributions in the tails of the pdf. Because of this result, power law tails (and thus the probability of occurrence of extreme events) have been hypothesized to have a relationship with non-Gaussian statistics.

3.6.3 Power law tails

The process of obtaining power-law tails from a stochastic climate model with multiplicative noise was first formulated by Sardeshmukh and Sura (2009) and later outlined by Sura and Perron (2010) and Sura (2011). One may assume that for a sufficiently large x and small noise amplitude, the Fokker-Planck equation given in Sardeshmukh and Sura (2009) may be simplified to give, in the notation of Sardeshmukh et al. (2015),

$$p_{\pm}(x) \approx \frac{1}{\mathcal{N}} [Ex + g]^{-2[1+(1/E^2)]} \exp\left(\pm \frac{\pi g}{E^2 b}\right), \quad (3.34)$$

where the $+/-$ sign corresponds to the respective positive/negative tail of the pdf.

From here, Sura (2011) shows that if one allows the probability density function, $p(x)$, to assume a power law distribution i.e., $p(x) \propto |x|^{-\alpha}$, then the power law decay (as named by Sardeshmukh et al. (2015)) requires

$$\alpha = 2 \left[1 + \left(\frac{1}{E^2} \right) \right]. \quad (3.35)$$

where α is the power law exponent.

Similarly, Sardeshmukh and Sura (2009) show that power law tails exist in the full solution of the Fokker-Planck equation, where a solution to the stationary pdf is achieved by approximating the probability density function as a power law. It should be noted that this analytical solution provides the same value of α for both the negative and positive power law tails i.e., the slope of the power law tails are symmetric about the full pdf. Symmetric power law slopes are not typically seen in observations (West (2012)).

From Equation (3.35), one would expect to see stronger power law tails when the multiplicative noise forcing is weaker (and vice versa). In other words, pdfs with stronger power law tails correspond to smaller values of α . Sura (2011) notes that the multiplicative noise is unable to approach zero, as it violates the assumption that x must be large. Instead, when the multiplicative noise approaches zero, the pdf simply becomes Gaussian.

CHAPTER 4

THE STOCHASTICALLY GENERATED SKEWED DISTRIBUTION

The one-dimensional pdf produced by the CAM noise stochastic model of the form given by Equation (3.29) was first written as the stochastically generated skewed (SGS) distribution by Sardeshmukh and Sura (2009). The work of Sardeshmukh et al. (2015) focused on the form and constraints of the SGS distribution given in Equation (3.29) and provided a methodology to estimate SGS parameters.

4.1 Properties of the SGS distribution

For clarity, the SGS probability distribution function of Equation (3.29), obtained from the CAM noise model of Sardeshmukh and Sura (2009) in the notation of Sardeshmukh et al. (2015), may be written as

$$p(x) = \frac{1}{\mathcal{N}} \left[(Ex + g)^2 + b^2 \right]^{-(1+(1/E^2))} \exp \left[\frac{2g}{E^2 b} \arctan \left(\frac{Ex + g}{b} \right) \right], \quad (4.1)$$

where the normalization constant, \mathcal{N} , was given previously in Equation (3.30) as

$$\mathcal{N} = \frac{2\pi\nu^{1/2} (2b)^{-(2\nu+1)} \Gamma(2\nu+1)}{\Gamma(\nu+1-iq/2) \Gamma(\nu+1+iq/2)}. \quad (4.2)$$

Here $\nu = 1/E^2$, $q = 2g\nu/b$ and Γ is the special gamma function. The SGS distribution parameters b and g have units that correspond to the standardized anomaly x , while E is unitless (Sardeshmukh et al. (2015)). Sardeshmukh et al. (2015) denotes the $(Ex + g)\eta_2$ as the CAM noise described in Sardeshmukh and Sura (2009), where $b\eta_1$ is the additive noise term.

The notation of Sardeshmukh et al. (2015) omits the damping term, λ , from the pdf which is included in some previous formulations of the SGS distribution, namely by Sardeshmukh and Sura (2009). The omission of the damping term does not affect the stationary SGS pdf, but it must be reintroduced later in computing the SGS Markov process as it constrains the temporal evolution of the time series. In practice, its omission merely reduces the number of SGS distribution parameters

to three. Sardeshmukh and Sura (2009) and Sardeshmukh et al. (2015) also note that the SGS pdf has a unique maximum at

$$x_{\max} = -\frac{Eg}{(1 + E^2)}. \quad (4.3)$$

The moments of the SGS distribution obtained in Equation (3.31) are given again here as

$$\begin{aligned} \langle x \rangle &= \mu = 0 \\ \langle x^2 \rangle &= \sigma^2 = \frac{g^2 + b^2}{2 - E^2} \\ \langle x^3 \rangle &= \text{skew} = \frac{2Eg}{\sigma(1 - E^2)} \\ \langle x^4 \rangle &= \text{kurt} = \frac{3}{2} \left[\frac{1 - E^2}{1 - (3/2)E^2} \right] \text{skew}^2 + \left[\frac{3E^2}{1 - (3/2)E^2} \right]. \end{aligned} \quad (4.4)$$

The Gaussian distribution itself is a member of the SGS class in the limit where $E \rightarrow 0$, which corresponds to a skewness and kurtosis of zero. Sardeshmukh et al. (2015) provides an analytical approximation of the SGS pdf in this “small- E ” limit, which may be written as

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp \left[\frac{-x^2}{2} + \frac{Sx}{6} (x^2 - 3) + \mathcal{O}(E^2) \right]. \quad (4.5)$$

Sardeshmukh et al. (2015) also point out that, in order for the kurtosis to exist, the parameter E must be constrained as

$$E < \sqrt{\frac{2}{n-1}} = \sqrt{2/3}, \quad (4.6)$$

where $n = 4$ represents the fourth moment. Sardeshmukh et al. (2015) also point out that E and b are positive, while g can be either positive or negative, such that the sign of g is the same as that of the skewness.

The power law behavior in the pdf tails described by Sardeshmukh and Sura (2009) may again be examined when $|Ex + g| \gg b$ (when the CAM noise is large), where again via Sardeshmukh et al. (2015) the pdfs for the positive (p_+) and negative (p_-) tails can be expressed respectively as

$$\begin{aligned} p_+(x) &\approx \frac{1}{\mathcal{N}} \exp\left(\frac{\pi g}{E^2 b}\right) [Ex + g]^{-2[1+(1/E^2)]} \\ p_-(x) &\approx \frac{1}{\mathcal{N}} \exp\left(\frac{-\pi g}{E^2 b}\right) [Ex + g]^{-2[1+(1/E^2)]}. \end{aligned} \quad (4.7)$$

Here, the slope of the power law distribution, $2[1 + (1/E^2)]$, is symmetric with respect to the positive and negative tails. Notably, Sardeshmukh et al. (2015) state that even after very long integrations of a Markov model, the power law decay begins tailward of ten standard deviations. The existence and symmetry of power law distributions in daily atmospheric data sets is a topic of current research.

4.2 Estimation of SGS parameters

The most straightforward means for calculating the parameters of the SGS distribution given the explicit forms of the SGS moments in Equation (4.4) is a method of moments. Indeed, Sardeshmukh et al. (2015) use a method of moments to estimate the parameters of the SGS distribution by rearranging the variance, skewness and kurtosis equations of Equation (4.4) and solving for E , b and g respectively. Parameter E is calculated first using values of skewness and kurtosis obtained from a sample population. Parameters b and g may then be computed using E and the sample variance and skewness:

$$\begin{aligned} E &= \sqrt{\frac{2}{3} \left[\frac{K - (3/2) S^2}{K - S^2 + 2} \right]} \\ b &= \sqrt{2\sigma^2 \left[1 - \frac{E^2}{2} - \frac{(1 - E^2)^2}{8E^2} S^2 \right]} \\ g &= S\sigma \left(\frac{1 - E^2}{2E} \right). \end{aligned} \tag{4.8}$$

The method of moments is a simple and computationally inexpensive means of estimating the parameters of the SGS distribution. However, the E , b and g parameters of SGS distribution are not the moments themselves e.g., like that of the Gaussian distribution. Using the method of moments to estimate model parameters themselves derived from the moments can lead to biases (Wilks (2006)). A more robust approach to estimating the parameters of the SGS distribution would be to use maximum likelihood estimation, though the implementation of this method is left to future research.

While it is straightforward to solve for the SGS parameters using the sample variance, skewness and kurtosis, values of the sample moments do not always meet the skewness-kurtosis inequality of Equation (3.32). Rewriting the skewness-kurtosis inequality in terms of Equation (4.4) gives

$$\text{kurt} = \frac{3}{2} \left[\frac{1 - E^2}{1 - (3/2) E^2} \right] \text{skew}^2 + \left[\frac{3E^2}{1 - (3/2) E^2} \right] > \frac{3}{2} \text{skew}^2. \tag{4.9}$$

Sura and Sardeshmukh (2008), Sardeshmukh and Sura (2009), Sardeshmukh et al. (2015) and others typically admit a small vertical offset (r , see Section 3.6.2 for discussion) when representing the skewness-kurtosis relationship in practice, but the method of moments has no such tolerance built in. As both terms in brackets left of the inequality cannot be less than zero, some observations tend to fall outside the inequality threshold. Most acutely, time series with negative kurtosis (of

which there are a significant amount in time series of daily averaged atmospheric data) are not able to be fit by the method of moments directly to an SGS distribution without violating the skewness-kurtosis relationship.

Sardeshmukh et al. (2015) also state that in order for E and g to be positive, another constraint must be applied to the estimate of E such that

$$E > \sqrt{\frac{\epsilon - 1}{\epsilon}}, \quad (4.10)$$

where $\epsilon = \sqrt{1 + (S^2/4)}$. Combining this constraint on E with Equation (4.6) yields

$$\sqrt{\frac{\epsilon - 1}{\epsilon}} < E < \sqrt{2/3}. \quad (4.11)$$

Equation (4.11) may be used to find the maximum magnitude of the skewness that can be represented by the SGS distribution while the kurtosis exists. In setting the lower and upper domain constraints of E equal to each other and solving for the skewness, the maximum magnitude of the skewness, S_{\max} , is found to be $\sqrt{32} \approx 5.65685$. Note that, similar to the skewness-kurtosis constraint of Equation (4.9), the skewness magnitude calculated from daily reanalysis data does not always fall below this bound.

To satisfy the model constraints imposed on fitting an SGS distribution to sample time series, Sardeshmukh et al. (2015) recommend that if the calculated value of E fails to meet the ϵ -constraint of Equation (4.10), the value of kurtosis should be increased and then used to recalculate E in Equation (4.8). This correction applied to the value of sample kurtosis is repeated until E satisfies Equation (4.10), at which point b and g may also be computed. Adjusting the kurtosis in this fashion deteriorates the fit of the SGS distribution with respect to the observations, but small changes in fourth moment have the smallest impact on the overall shape of the distribution.

As such, there are three ways that the SGS model constraints may be violated (and the calculation of parameter E is affected) when using the method of moments to fit an SGS distribution from sample data:

1. The magnitude of the sample skewness is greater than $\sqrt{32}$, i.e., $|\text{skew}| > \sqrt{32}$
2. $E < 0$, such that the ϵ -constraint is unmet
3. $E > \sqrt{2/3}$, such that the kurtosis does not exist

However, rather than modifying the values of the skewness or increasing kurtosis for the purposes of calculating the SGS parameters as in Sardeshmukh et al. (2015), a value of E is instead assigned directly and equivalently that meets the above constraints. Namely:

- If the sample skewness is larger than $\sqrt{32}$, E is set equal to $\sqrt{2/3}$, the maximum model value of E allowed when the kurtosis exists
- If $E < 0$, the value of E is then set to the minimum value allowed by the ϵ -constraint of Equation (4.10), itself a function of the sample skewness
- If the value of E calculated is larger than $\sqrt{2/3}$, E is simply set to $\sqrt{2/3}$

The task of Chapter 6 is to show that the SGS parameters estimated by the method of moments ultimately produce a satisfactory SGS distribution, in spite of some constraint violations, that is representative of daily-averaged atmospheric reanalysis data.

4.3 The SGS Markov process

Sardeshmukh et al. (2015) write the CAM noise Markov process model, first introduced by Sardeshmukh and Sura (2009), capable of creating synthetic time series of SGS distributions, as

$$dx = - \left[\left(1 + \frac{1}{2}E^2 \right) x + \frac{1}{2}Eg \right] \lambda dt + [b\eta_1 + (Ex + g)\eta_2] \sqrt{\lambda} dt. \quad (4.12)$$

New time series values may be obtained by iterating the model $x(t)$ by an increment dx . Again, η_1 and η_2 are delta correlated Gaussian white-noise terms with zero mean and unit variance. λ is equal to $1/\tau_c$ where τ_c is the decorrelation timescale¹, calculated by estimating the autocorrelation function of the observations which informed the SGS distribution at increasing lag times, which serves as a time damping and standardization.

By estimating SGS parameters E , b and g , one can numerically integrate Equation (4.12) using a stochastic solver. This process is undertaken by the `SGSDist.jl` Julia package and described in Chapter 5.

¹The decorrelation timescale is the time lag at which the time series is no longer correlated with itself, where the autocorrelation function $\rho(\tau) = e^{-\lambda\tau}$.

CHAPTER 5

THE SGSDIST.JL PACKAGE

A Julia package, `SGSDist.jl`, has been developed to make the creation, estimation and evaluation of SGS distributions efficient and effective to use. A package such as `SGSDist.jl` is needed in order to satisfy, with computational efficiency:

- Estimation of distribution parameters E , b and g
- Calculation of the SGS pdf, CDF and other statistical quantities for arrays of random variables
- Drawing of random numbers from an SGS distribution
- Creation of synthetic time series with statistics informed from a particular SGS distribution

This package is fundamental to the results of Chapter 6, where it is used extensively to evaluate the non-Gaussianity of 20CRv2c atmospheric variables and construct statistical significance tests.

`SGSDist.jl` has been publicly released under the MIT license. The source code can be viewed at the package GitHub page at <https://github.com/brwst/SGSDist.jl>. More discussion and examples on the functionality of the `SGSDist.jl` module may be found in the package documentation located at <https://brwst.github.io/SGSDist.jl>.

5.1 Installation

The Julia programming language is a dynamically-typed language with similar syntax to Python with the performance of Fortran or C. It is developed for use as a scientific programming language, aiming to become an open-source alternative to MATLAB or Mathematica. More information on Julia, including how to download and run it, can be found in Bezanson et al. (2017) and at the project's documentation page, <https://docs.julialang.org/>.

Assuming a working Julia installation, one may install the `SGSDist.jl` package by issuing the following at the Julia REPL¹ prompt:

```
julia> using Pkg; Pkg.add("SGSDist.jl")
```

¹The Julia read-eval-print-loop interactive environment.

or, with the Julia package manager,

```
(v1.3) pkg> add SGSDist.jl
```

One may then load the `SGSDist.jl` package with:

```
julia> using SGSDist
```

The rest of the chapter is devoted to describing a basic analysis workflow using the `SGSDist.jl` package. For brevity, REPL spacing is condensed and command output is sometimes truncated and replaced with ellipsis.

5.2 Creating an SGS distribution

After installing `SGSDist.jl`, using the package is straightforward. Here, as an example, one may create an SGS distribution with the same parameters as the ideal SGS distribution discussed in Sardeshmukh et al. (2015). This distribution is fit using the method of moments `fit()` function with parameters of unit variance and skewness but a kurtosis of 5.

```
julia> using SGSDist
julia> d = fit(SGS, 1.0, 1.0, 5.0)
SGS{Float64}(E=0.6236095644623235, b=1.1709106481844573, g=0.48997894350611143)
```

Here, `d` is the SGS distribution that has been fitted to the moments specified with parameters `E`, `g` and `b`. As mentioned in Section 4.2, the method of moments sometimes requires the value of parameter `E` to be artificially set in order to satisfy the SGS model constraints. If this occurs, a warning will be displayed indicating the specific constraint violation and action taken.

5.3 Fitting an SGS distribution from data

An SGS distribution may also be fit from an array of data. For example, one may use the method of moments described in Section 4.2 to fit a sample reanalysis time series of standardized temperature anomalies, provided by the function `air()`, to an SGS distribution by issuing:

```
julia> d = fit(SGS, air())
SGS{Float64}(E=0.36623331624359673, b=0.7133796999284913, g=-1.1648873051087967)
```

In this fashion, a multidimensional array of SGS distributions can be produced by applying the method of moments to time series of global reanalysis datasets.

The first four moments of the SGS distribution, written in Equation (4.4), may be calculated from a prescribed SGS distribution.

```
julia> mean(d)
0.0
julia> var(d)
0.9999999312924792
julia> skewness(d)
-0.9854112036580926
julia> kurtosis(d)
2.082561683550699
```

This output serves as a reminder that any input time series to be fitted to an SGS distribution must be composed of standardized anomalies i.e., mean of zero and unit variance.

5.4 Calculating the SGS pdf

One may calculate the SGS pdf of the form specified in Equation (3.29) for a single variable or an array of variables. Using the SGS distribution `d` that was fit from the sample temperature anomalies above, one may compute the pdf by issuing:

```
julia> pdf(d, 0.4)
0.45253005501315047
```

By loading the `StatsPlots.jl` package, one may plot the estimated SGS pdf and compare it to a step histogram of the observations. This output is shown in Figure 5.1.

```
julia> using StatsPlots
julia> x = collect(-5.0:0.1:5.0);
julia> plot(x, pdf.(d, x), label="SGS pdf of air()", xlabel="T", ylabel="p(T)")
julia> stephist!(air(), norm=:pdf, label="air()")
```

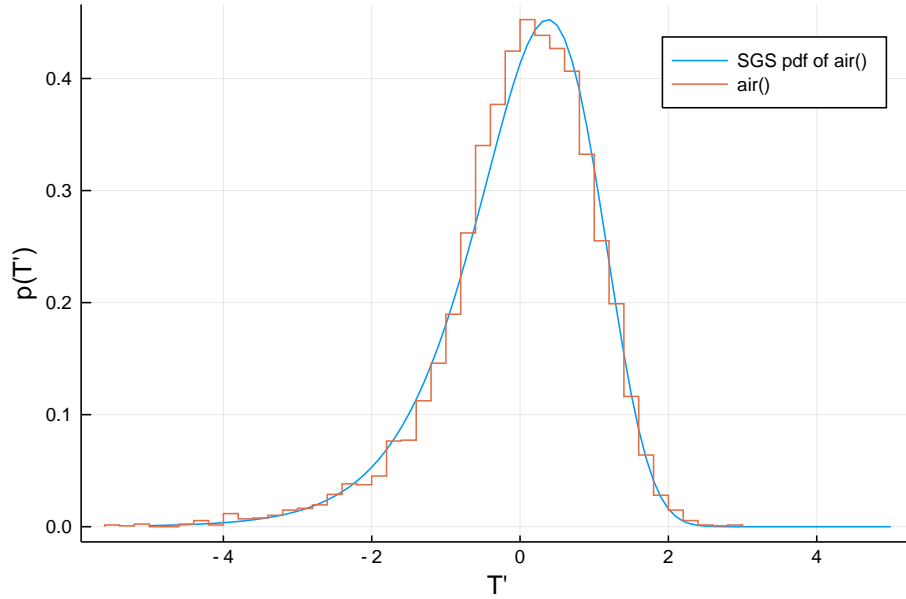


Figure 5.1: Plot of the example air temperature SGS pdf, showing the SGS pdf ($\sigma^2 = 1$, skew = 1, kurt = 5) (blue) in comparison with the step histogram of the sample air temperature data (red).

Note that the `SGS pdf()` function will compute the “small E ” pdf approximation for $E < 0.08$. This value of was chosen based on results from testing the stability of the pdf at low- E values. As such, the convergence of the SGS pdf to the Gaussian distribution when $E \rightarrow 0$ is reflected in Figure 5.2, which may be produced via:

```
julia> using Distributions                # needed for Normal()
julia> d = SGS(0.05, 0.9, -1.11)         # a "small-E" SGS distribution
SGS{Float64}(E=0.05, b=0.9, g=-1.11)
julia> x = collect(-5.0:0.1:5.0)
julia> plot(x, pdf.(d, x), label="SGS", xlabel="x", ylabel="p(x)")
julia> plot!(x, pdf.(Normal(), x), label="Std. Norm.")
```

5.5 Calculating the SGS CDF

While an analytical form of the SGS cumulative distribution function (CDF) has not yet been derived, the CDF of a particular SGS distribution may be approximated by numerical integration.

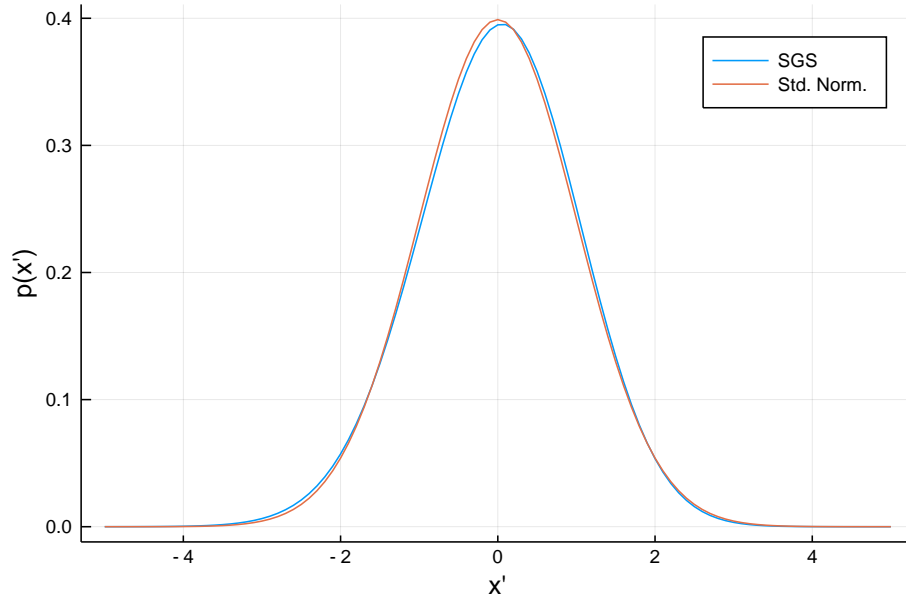


Figure 5.2: Plot of the "small E" SGS pdf, showing the SGS pdf (blue) in comparison with the standard normal pdf (red) when $E \rightarrow 0$.

Here, an adaptive one-dimensional Gaussian quadrature scheme is implemented using the Julia `QuadGK` package², where the SGS pdf is integrated from -30 standard deviations to the desired upper bound, x . While the `QuadGK` package has support for integrating over the domain $(-\infty, \infty)$, a lower bound of `-Inf` occasionally produced numerical instabilities in some near-Gaussian distributions.

The CDF of the example SGS distribution `d` is shown in Figure 5.3 and can be computed by simply issuing `cdf(d, x)`:

```
julia> x = collect(-5.0:0.1:5.0);
julia> plot(x, cdf.(d, x), label="CDF of air()", xlabel="T", ylabel="F(T)")
```

5.6 Random number generation

A random number or an array of random numbers may be drawn from an SGS distribution `d` by invoking `rand()`.

²The source code and links to the documentation for the `QuadGK` package are located at <https://github.com/JuliaMath/QuadGK.jl>.

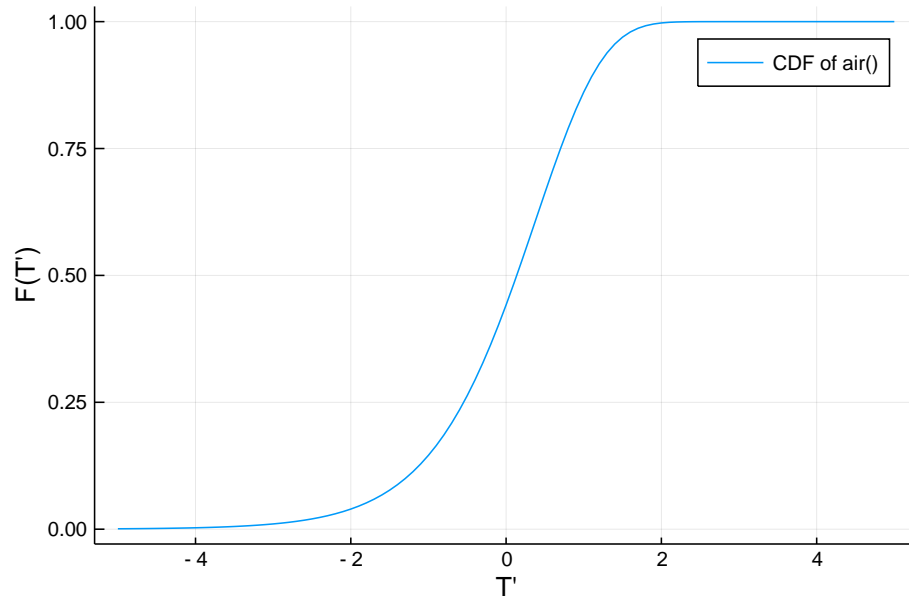


Figure 5.3: Plot of the SGS CDF (blue) obtained by estimating the SGS distribution from example air temperature data.

```
julia> rand(d, 3)
3-element Array{Float64,1}:
-0.6658576564565255
-1.8058720924121987
 0.08814146377788976
```

This is achieved through an inverse transform sampling method, which maps uniform random number generation to a non-uniform random number. Essentially, this method relies on the fact that the CDF of the non-uniform distribution (here the SGS distribution) is always uniform and bounded on $[0, 1]$. Conveniently, this is also the domain of the uniform random number generator. Therefore, by inverting the CDF of the SGS distribution and applying a random variate generated by the uniform random number generator, one obtains an SGS random variate that is distributed with respect to the CDF of the SGS distribution. See Wilks (2006) for more discussion on inverse transform sampling.

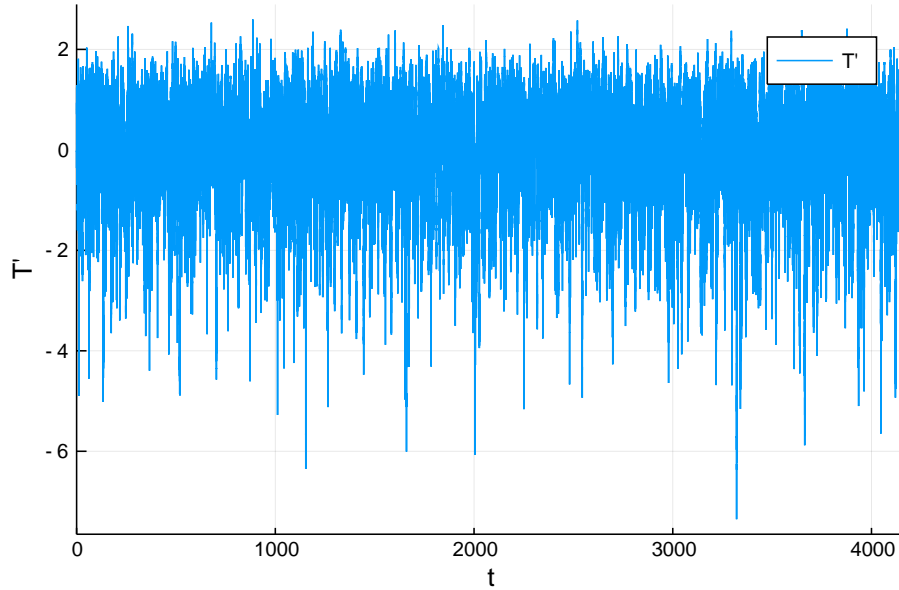


Figure 5.4: Plot of the CAM1D Markov process time series with default arguments ($dt=1/24$, $\lambda=1$), where the statistics of the time series are informed from an SGS distribution with $\sigma^2 = 1$, skew = 1, kurt = 5.

5.7 Markov processes

One may create a time series via Markov process that represents the statistics of a given SGS distribution. To do so, one must first specify the candidate SGS distribution. Here, a sample SGS distribution is again created with a variance and skewness of 1 and a kurtosis of 5:

```
julia> d = fit(SGS, 1, 1, 5)
SGS{Float64}(E=0.6236095644623235, b=1.1709106481844573, g=0.48997894350611143)
```

The one-dimensional correlated additive and multiplicative (CAM) noise model described in Sardeshmukh and Sura (2009) and written in Equation (4.12) can be used to produce a Markov process time series with the statistics of the SGS distribution `d` by invoking the `CAM1D()` function.

```
julia> n = 100000; # length of time series
julia> plot(CAM1D(d, n), label="T", ylabel="T")
```

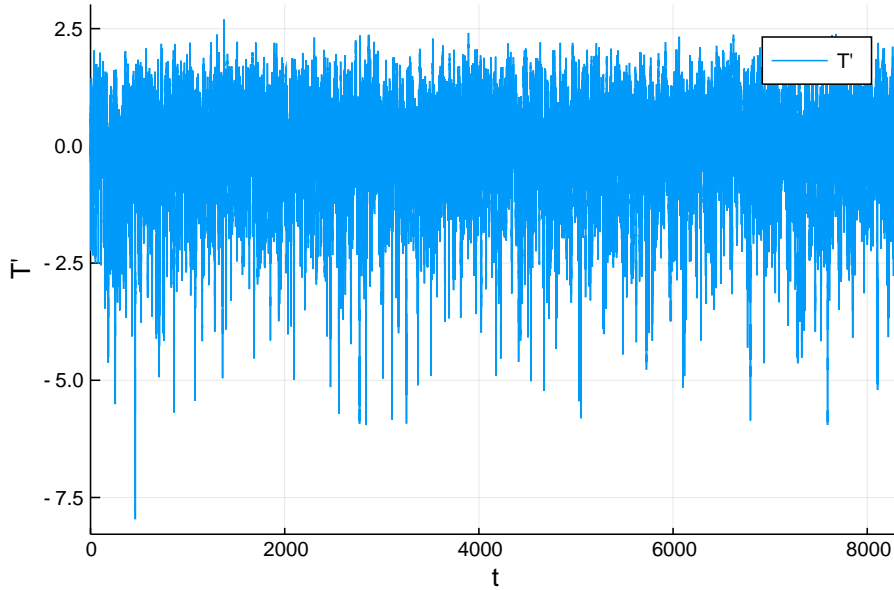


Figure 5.5: Plot of the CAM1D Markov process time series with $dt=1/12$, $\lambda=0.5$ and a seed provided to the random number generator. The statistics of the time series are informed from an SGS distribution with $\sigma^2 = 1$, skew = 1, kurt = 5.

The Markov process detailed in Equation (4.12) is implemented by numerically iterating a stochastic differential equation solver using Heun’s method.³ By default, the `CAM1D()` function assumes an hourly timestep argument ($dt=1/24$) and a damping factor of 1.0.

One may specify a different value for the time step or damping term. Moreover, a seed may be assigned in order to reproduce a time series over subsequent runs.

```
julia> plot(CAM1D(d, n, dt=1/12, lambda=0.5, seed=42), label="T", ylabel="T")
```

Plotting a step histogram of the generated time series in conjunction with the fitted SGS pdf (here in Figure 5.6) allows an assessment of the Markov process solver with respect to the distribution statistics.

```
julia> sol = CAM1D(d, n, dt=1/12, lambda=0.5, seed=42);
julia> T = collect(map(x -> x[1], sol.u));
```

³This is achieved by the `SGSDist.jl` package in two ways. By default, the `EulerHeun()` algorithm is invoked from the `DifferentialEquations` Julia package, but a fourth-order Runge-Kutta method has also been implemented for testing purposes.

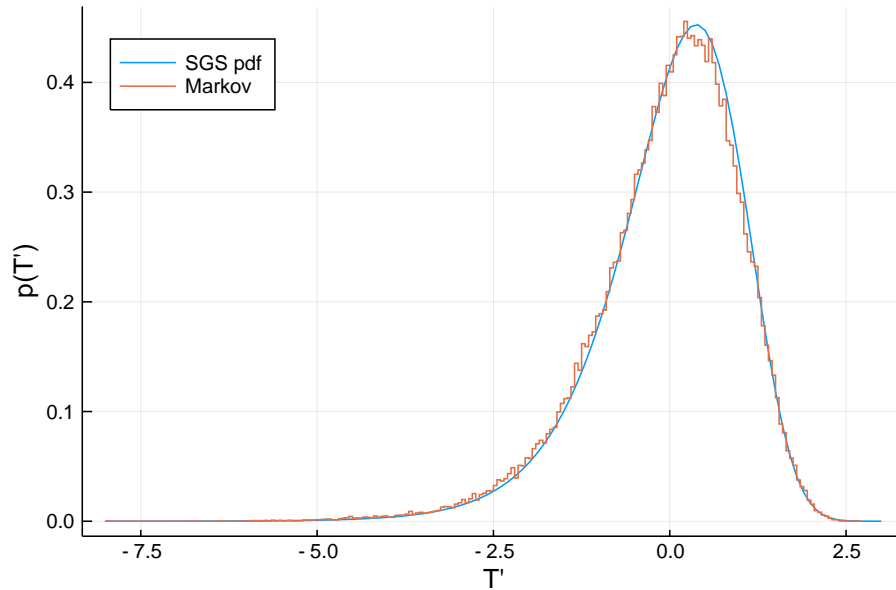


Figure 5.6: Step histogram (blue) of the previous CAM1D Markov process time series with $dt=1/12$, $\lambda=0.5$ along with the pdf (red) of the SGS distribution used to produce the Markov process time series.

```
julia> plot(x, pdf.(d, x), label="SGS pdf", legend=:topleft, xlabel="T", ylabel="p(T)")
julia> stephist!(T, norm=:pdf, label="Markov")
```

In addition to the CAM noise model, it is also possible to create a Markov process time series using the Hasselmann climate model, discussed in 3.5 and described in Hasselmann (1976). The Hasselmann Markov process is an AR(1) process which produces a time series that is normally distributed.

```
julia> n = 1000; # length of time series
julia> plot(Hasselmann1D(n))
```

As in the case of the CAM noise model, one may also specify the time step, damping terms and the seed used for the random number generation.

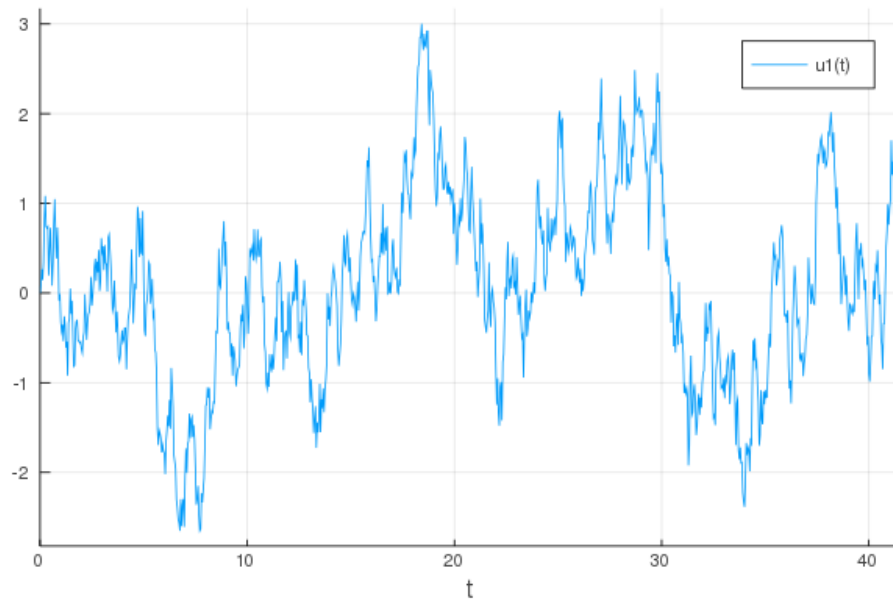


Figure 5.7: Plot of the `Hasselmann1D` Markov process time series with default arguments (`dt=1/24`, `lambda=1`), where the statistics of the time series are informed from a standard normal distribution.

CHAPTER 6

STATISTICAL EVALUATION OF THE SGS DISTRIBUTION

This chapter assesses whether the SGS distribution can plausibly represent the empirical distributions of atmospheric variables from the 20CRv2c dataset. Here, persistent wintertime (DJF) standardized anomalies from 1947-2014, referred to here as the “modern era” of the 20CRv2c dataset, are examined to establish the statistical test methodology and to provide a baseline for further comparisons in Chapter 7. Chapter 8 will investigate DJF standardized anomalies from 1880-1947 (the “historical era”) and draw comparisons with the modern era.

6.1 Fitting an SGS distribution to a time series

The methodology for evaluating the validity of the SGS distribution using reanalysis data is demonstrated first in the one-dimensional case by selecting two time series of 20CRv2c 950 hPa air temperature anomalies, one located in a region of significant near-Gaussianity, the other in a region of non-Gaussianity. Figure 6.1 provides a global view of 1947-2014 DJF skewness and kurtosis obtained from the 20CRv2c air temperature anomalies, from which the two points are selected.

Point A, located east of Hawaii (referred to hereafter as East Hawaii) at 20°N , 138°W , is in a region of near-Gaussianity due to the near-zero magnitudes of skewness and kurtosis. Point B, located near Vancouver, British Columbia (referred to hereafter as Vancouver), is a non-Gaussian time series, located at 50°N , 124°W in a region of negative skewness and positive kurtosis. An SGS distribution is fitted to both standardized time series by using the method of moments described in Section 4.2 and the procedure outlined in Section 5.2. Table 6.1 describes the point locations, statistics and fitted SGS distribution parameters of the two locations being tested. Note that East Hawaii has a small value of E , which corresponds closely to the Gaussian distribution as discussed in Section 4.1 where $E \rightarrow 0$ as the SGS pdf converges to the Gaussian distribution. East Hawaii also has a negative value of g , corresponding to the negative value of skewness at that location.

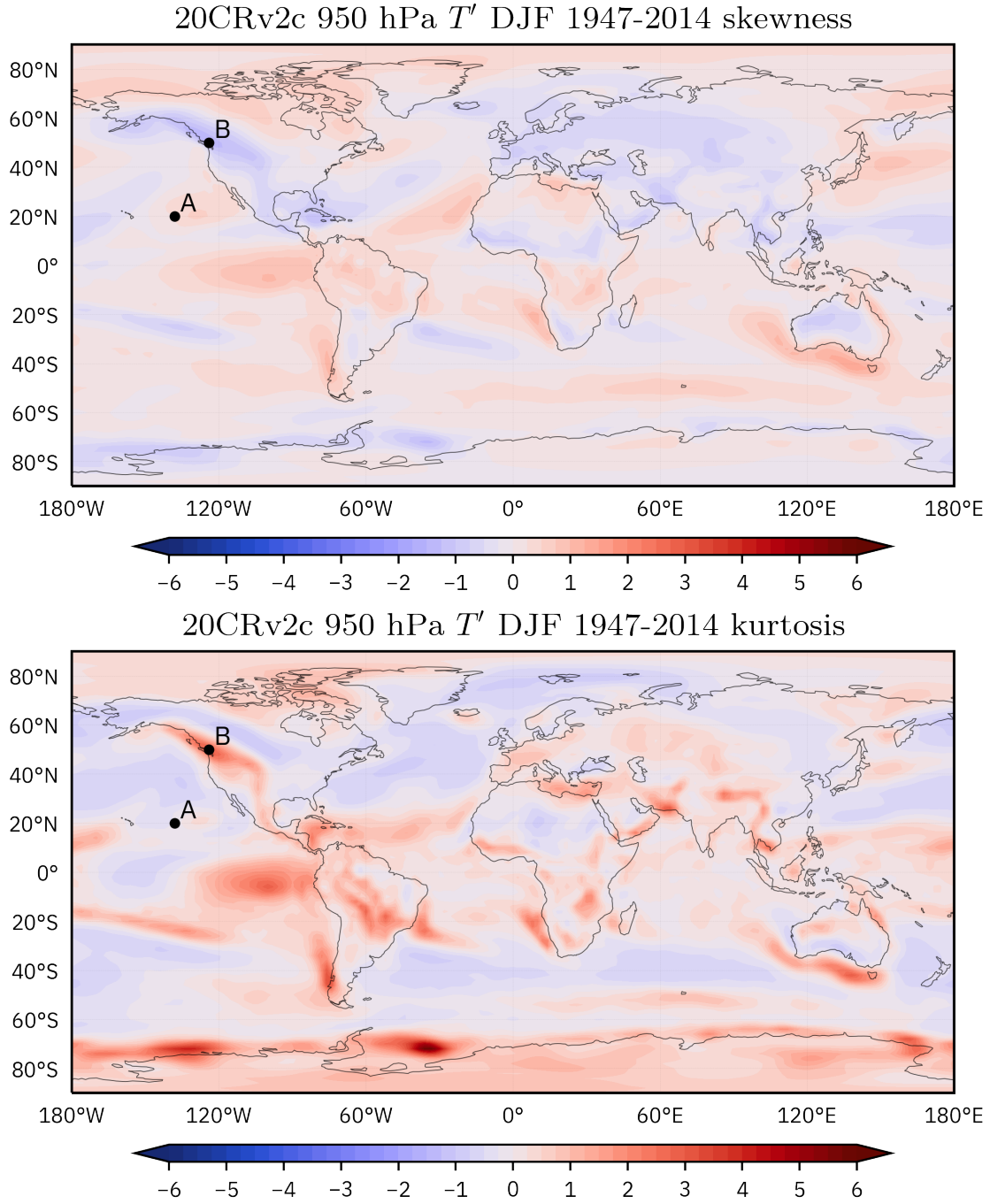


Figure 6.1: Maps of skewness and kurtosis of standardized DJF 950 hPa 20CRv2c air temperature anomalies, 1947-2014. Point A (East Hawaii, 20°N, 138°W) and point B (Vancouver, 50°N, 124°W) are delineated on the map.

Similarly, E is larger at the more non-Gaussian Vancouver location and has a positive value of g corresponding to the positive kurtosis at that grid point.

Table 6.1: Locations, higher moments and SGS parameters E , b and g of near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) 1947-2014 DJF 950 hPa air temperature anomaly time series.

Point	Location	Skewness	Kurtosis	E	b	g
A	East Hawaii	0.268	-0.187	0.104	0.607	1.273
B	Vancouver	-1.416	2.968	0.439	0.331	-1.303

The two resultant SGS pdfs are shown in Figure 6.2, calculated using Equation (3.29), where the red curve denotes the fitted SGS distribution, while the blue curve represents the stepwise histogram of the air temperature anomalies. For reference, the standard normal distribution is included in gray.

While the numerical tools described in Section 5.2 can fit an SGS distribution to the time series and provide an estimate for the SGS parameters E , b and g , they do not tell us if the estimated SGS fit is representative of the data. In other words, given a time series of standardized anomalies, one may always fit an SGS distribution to the data whether it is a good fit or not.

However, there are also an infinite number of pdfs that may be constructed that fit the empirical data “better” than that of the SGS distribution. A “perfect” theoretical representation of the observations will still be subject to errors from the statistical fluctuations inherent in the observations. As such, the goal is to quantitatively test the SGS distribution against variables in the 20CRv2c dataset to establish 1) whether the SGS distribution is a plausible fit to the data and 2) that the SGS distribution is a better representation of the data than the Gaussian distribution.

6.2 Visual comparison with Q-Q plots

For a start, this comparison may be done visually using a quantile-quantile (Q-Q) plot, which is useful for graphically comparing the empirical distribution of observations with a reference distribution. The observations of a random variable x are plotted against the quantile function of the distribution of interest, producing a scatter plot of values. A Q-Q plot with a distribution that fits the data perfectly would have all points align on the $y = x$ diagonal (Wilks (2006)).

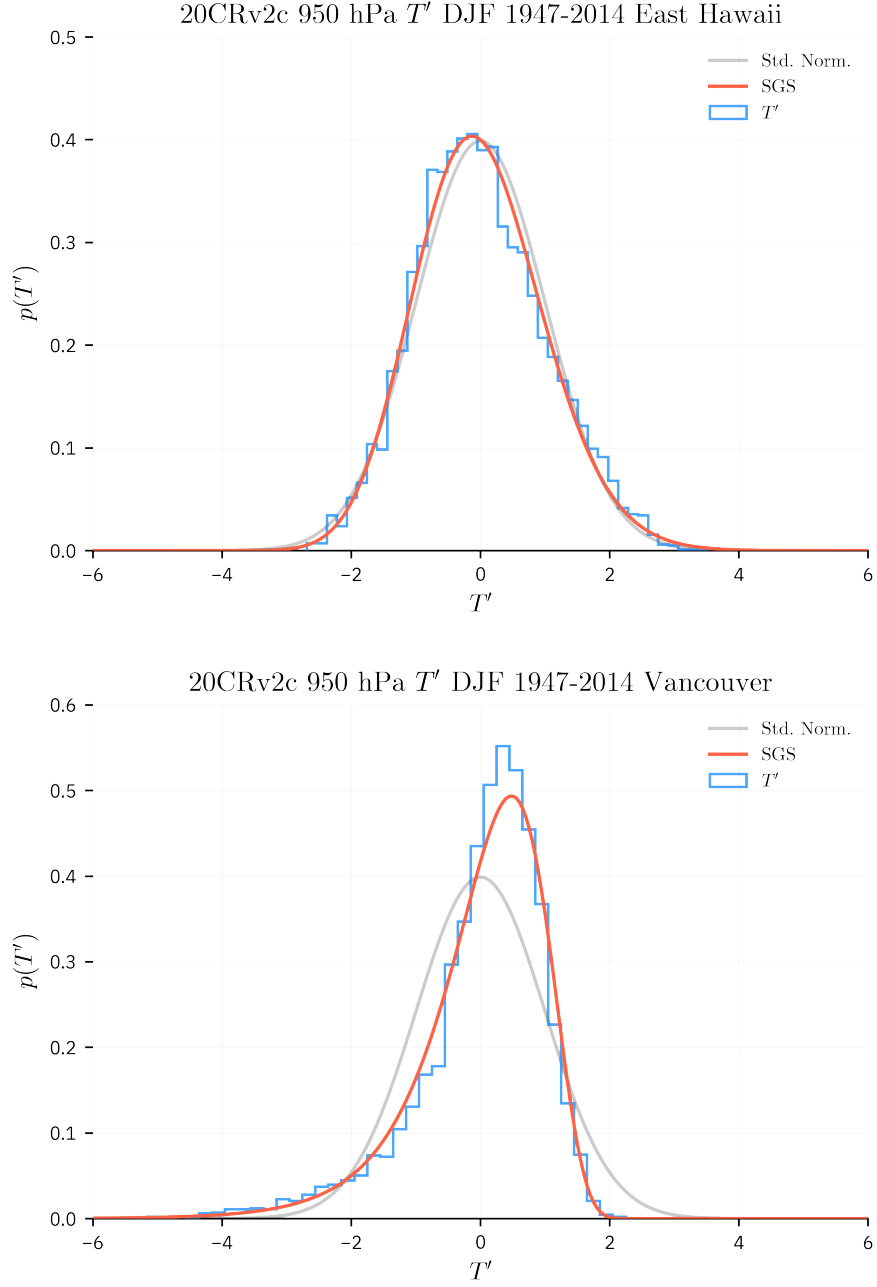


Figure 6.2: SGS pdfs (red) and stepwise histograms (blue) of the DJF 950 hPa 20CRv2c air temperature anomalies (1947-2014) located at near-Gaussian East Hawaii (top, 20°N, 138°W) and at non-Gaussian Vancouver (bottom, 50°N, 124°W). The standard normal distribution (gray) is provided for reference.

Figure 6.3 shows Q-Q plots of the 950 hPa air temperature anomalies (T') at East Hawaii against both the standard normal and estimated SGS distribution, while Figure 6.4 shows the Q-Q plots of Vancouver with respect to the same two distributions.

From Figure 6.3, the distribution of near-Gaussian air temperature anomalies at East Hawaii seem to generally agree with both the standard normal distribution and the estimated SGS distribution. However, it appears that the SGS distribution fits the data slightly better in the tails of the air temperature anomalies, though this difference is marginal. Figure 6.4 details the Q-Q plots obtained from the non-Gaussian air temperature anomalies at Vancouver, with more conclusive results. It is evident that the estimated SGS fit does a better job in representing the data in the center and positive tail of the distribution, as the air temperature anomaly quantiles there follow the $y = x$ line closely. However, the estimated SGS negative tail is still too strong with respect to the observations. The comparison with the standard normal distribution at Vancouver shows that a standard normal distribution predicts tails that are too weak and underestimates the peakedness of the data. Moreover, the arcing shown in the standard normal quantile Q-Q plot demonstrates the negative skewness of the air temperature anomalies and that the observations at Vancouver are incompatible with a standard normal distribution.

While Q-Q plots lend themselves to useful graphical analysis, a hypothesis testing methodology must be more than just visual inspection or other qualitative measures. Specifically, it is important to distinguish between the natural statistical fluctuations intrinsic to atmospheric data and that of data poorly distributed with respect to an SGS distribution. To address this, a bootstrapping goodness-of-fit test is implemented which produces a quantitative value of the plausibility of the distribution fit.

6.3 A goodness-of-fit test using the Kolmogorov-Smirnov statistic

After obtaining a time series of standardized anomalies and an SGS distribution fitted from the same anomalies, it is imperative to test whether the estimated SGS distribution is plausible given the data from which it was fit. To undertake this task, a goodness-of-fit methodology described by Clauset et al. (2009), who tests the fit of power law distributions to empirical data using a bootstrapping method and the Kolmogorov-Smirnov statistic, is implemented.

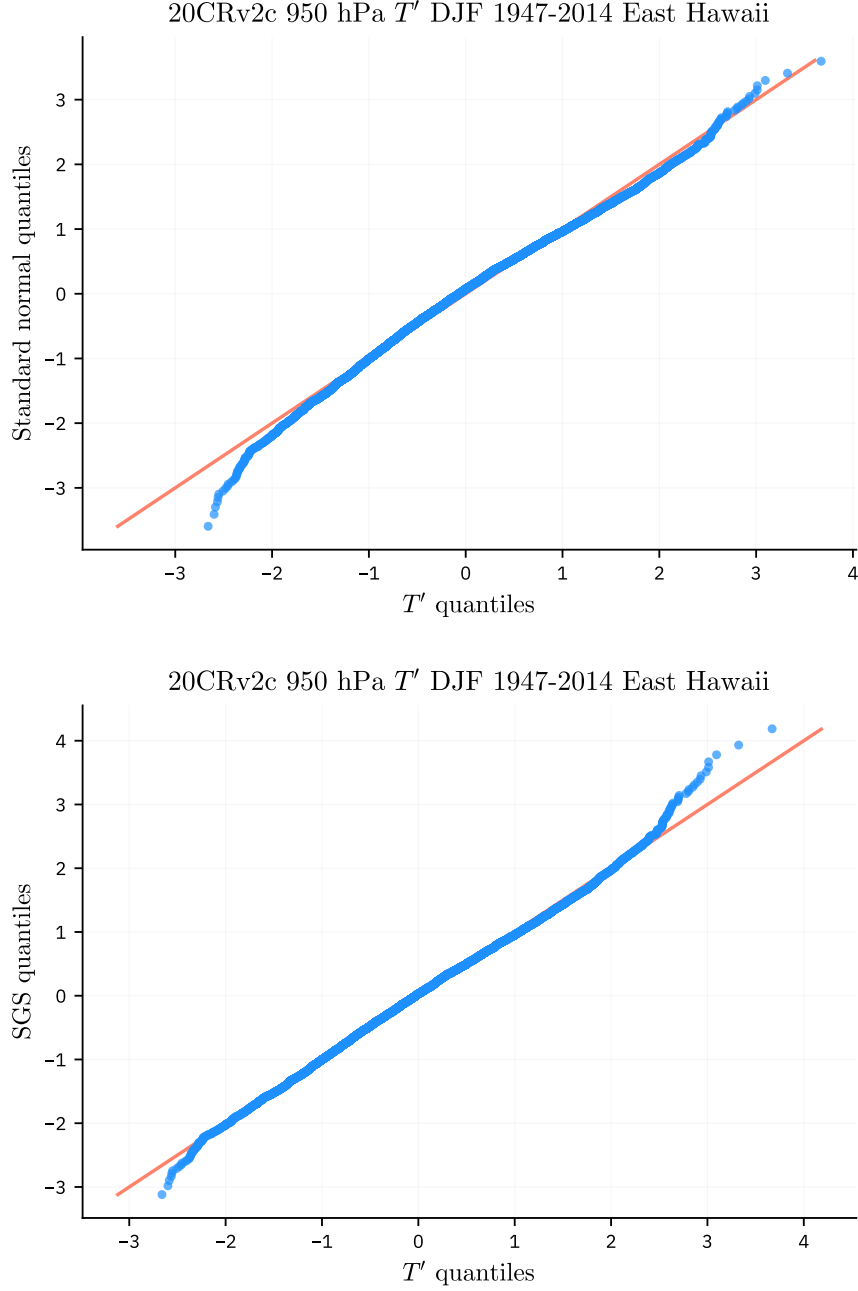


Figure 6.3: Q-Q plots of the DJF 950 hPa 20CRv2c air temperature anomaly (1947-2014) quantiles (blue dots) located at near-Gaussian East Hawaii (20°N , 138°W) compared against the quantiles of the standard normal distribution (top) and estimated SGS distribution (bottom). If the two distributions are similar, the air temperature quantiles will lie on the red $y = x$ line.

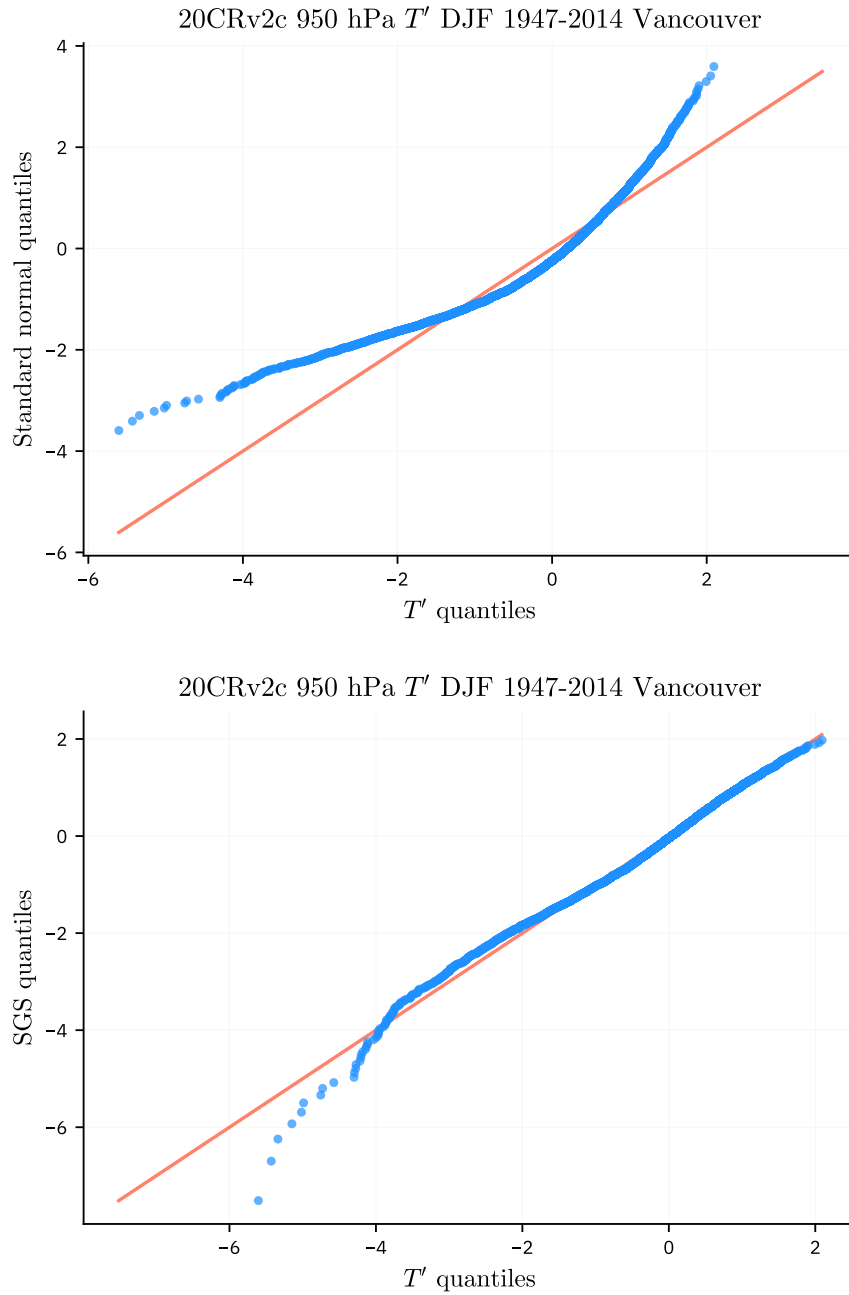


Figure 6.4: Q-Q plots of the DJF 950 hPa 20CRv2c air temperature anomaly (1947-2014) quantiles (blue dots) located at non-Gaussian Vancouver (50°N , 124°W) compared against the quantiles of the standard normal distribution (top) and estimated SGS distribution (bottom). If the two distributions are similar, the air temperature quantiles will lie on the red $y = x$ line.

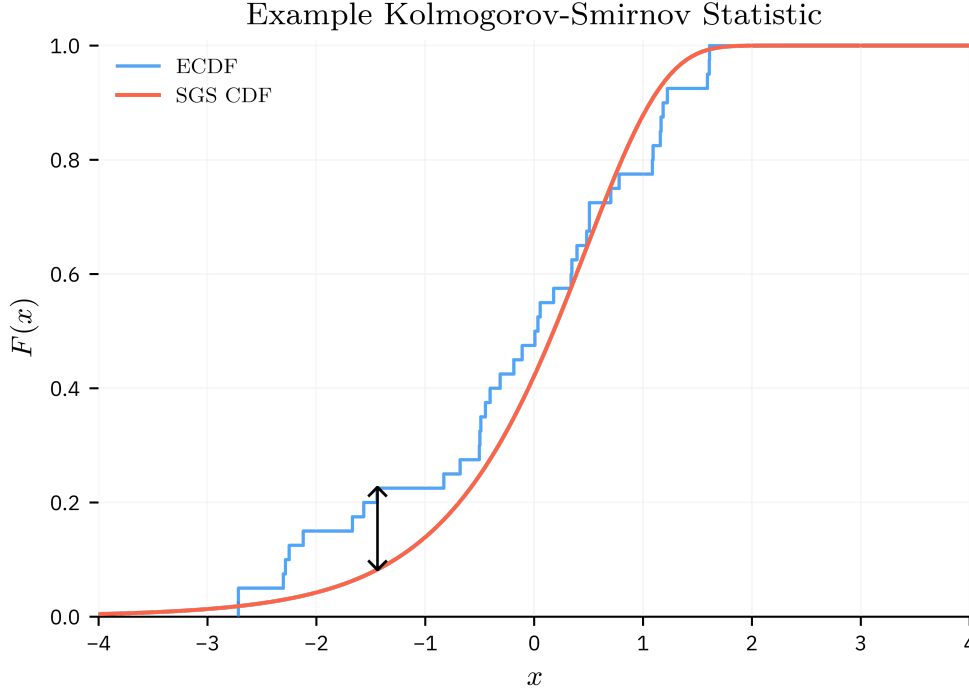


Figure 6.5: Two example cumulative distribution functions (CDF or $F(x)$): an empirical cumulative distribution function (ECDF, blue) assembled from the example time series x and the estimated SGS cumulative distribution function (red) obtained by fitting an SGS distribution to x . The Kolmogorov-Smirnov statistic, D , is the maximum difference between the two CDFs, shown by the black arrow.

The Kolmogorov-Smirnov statistic¹ is a statistical distance that is computed between the empirical cumulative distribution function (ECDF) of the data and the CDF of the estimated SGS distribution. The KS statistic may be written as

$$D = \max_x |F(x) - P(x)|, \quad (6.1)$$

which computes the maximum distance, D , between the corresponding points of ECDF ($F(x)$) and the CDF of the estimated SGS distribution ($P(x)$). A graphical example of this distance difference is shown in Figure 6.5.

The bootstrapping methodology may be summarized in a few steps. First, the Kolmogorov-Smirnov statistic is computed for the estimated SGS distribution, relative to the ECDF constructed

¹Also referred to as the K-S or KS statistic. In practice, another statistical test, such as the Anderson-Darling test, could be substituted here for the Kolmogorov-Smirnov statistic.

from the observations. Then, many pseudorandom datasets drawn from the estimated SGS distribution are produced, then fit themselves to an SGS distribution. The new SGS distribution obtained from the pseudorandom data is then used to compute a new KS statistic. The KS statistic comparing each pseudorandom SGS distribution with the pseudorandom data is tabulated and compared to the KS statistic of the original SGS fit to the original data.

When comparing the KS statistics of the pseudorandom data against the original dataset, one may say that if the original data has a larger KS statistic than the pseudorandom KS statistic, then the original SGS distribution is a worse fit than random statistical fluctuations and is not a plausible fit to the observations. However, if the original data has a smaller KS statistic than that of the pseudorandom data, then the fit of the original SGS distribution is plausible. Of course, it is possible that a good fit may sometimes lose out to statistical fluctuations. Clauset et al. (2009) remind us however that as the length of the observations used in the initial fit increases, the chance that a non-SGS derived pseudorandom dataset will produce an SGS distribution decreases.² Performing many bootstrap iterations produces a fraction of instances where the pseudorandom KS statistic is larger than the KS statistic derived from the original data and fit. This fraction yields a p-value, which can be used to quantify the plausibility of the original SGS fit.

Sometimes p-value thresholds are set to delineate “significant” and “insignificant” statistical fits. This practice, though common, can lead to pitfalls when interpreting the results. Often a p-value significance threshold of 0.1 (or a less stringent 0.05) is used to distinguish between a significant and insignificant statistical model. In other words, if the p-value falls below the threshold the statistical model under study is deemed an insignificant fit, but if the p-value sits above the threshold then the model is a plausible fit to the data. However, this introduces arbitrary significance cutoffs that may or may not be stringent enough to filter out weakly plausible distributions. Moreover, the presence of an arbitrary threshold gives the impression of a binary significance-insignificance choice, when in fact the magnitude of the p-value is sometimes ambiguous and itself subject to statistical fluctuations and therefore should be taken with caution. For these reasons, an arbitrary p-value cutoff is not introduced for the purposes of assigning significance. Instead, following Clauset et al. (2009), if the p-value is large (at least greater than 0.1) and approaches 1, then the estimated fit will be deemed plausible and differences between the observations and the fit may be attributed

²Clauset et al. (2009) notes that this is the reason why large sample time series are desired in the first place.

to statistical fluctuations. If the p-value is close to zero, then the estimated fit is unlikely to be representative of the observations. At times it may be convenient to introduce an arbitrary threshold, such as a p-value of 0.1, to serve as a benchmark of comparison or plausibility. However, this threshold will not be used to assign a level of significance.

The bootstrapping method of Clauset et al. (2009) is the preferred “goodness-of-fit” test for autocorrelated data, like that of the time series of variables obtained from the 20CRv2c dataset. An autocorrelated time series is often not applicable with classic statistical tests that assume the input sample data is independent and identically distributed (i.i.d.). In fact, the Kolmogorov-Smirnov statistic itself offers a simple way to compute a p-value based on the original model fit to the observations, as the distribution of Kolmogorov-Smirnov distances themselves follow the Kolmogorov distribution if the data is independently drawn from the same probability distribution. If the sample data is independent, one can simply map the KS statistic to the appropriate quantile of the Kolmogorov distribution to calculate a p-value. However, due to the time-dependence of the reanalysis data (as well as determining the SGS distribution by fitting to the data), a bootstrapping technique must be used to achieve a p-value.

6.4 Treating temporal correlations in reanalysis data

The temporal autocorrelation of the reanalysis data plays a significant role in the bootstrapping goodness-of-fit test. When generating realizations of pseudorandom time series drawn from the estimated SGS distribution, the length of the pseudorandom time series must be normalized by the decorrelation time of the variable of interest. According to Storch and Zwiers (2001), the decorrelation time is a “statistical measure that compares the information content of correlated observations with that of uncorrelated observations.” In other words, the decorrelation time is a scaling of the length of the original correlated time series in order to represent the same amount of information in an uncorrelated time series. In practice, the decorrelation time is lag time where the estimated autocorrelation of the time series of interest becomes negligible.

Scaling the uncorrelated pseudorandom time series appropriately for use in a goodness-of-fit test requires the calculation of the decorrelation time of each reanalysis time series. This decorrelation time, or lag time, can then be used to appropriately set the length of the pseudorandom dataset for statistical testing. Obtaining a reasonable decorrelation time from the reanalysis time series is

very important for the bootstrapping goodness-of-fit test. If the decorrelation time is too long, the pseudorandom dataset will be too short. This means the CDF used to calculate the KS statistic may be too coarse relative to the ECDF and may be subject to enhanced statistical fluctuations. If the decorrelation time is too short, the pseudorandom dataset will be too long and the CDF used to calculate the KS statistic may be too smooth relative to the ECDF and converge too quickly to the “true” CDF.

In order to compute a decorrelation time (τ) from a time series of reanalysis data, the “large-lag standard error” method described by Box et al. (1994) and Bartlett (1946) was initially used. This method attempts to compute the time lag at which the autocorrelation of the data is practically zero. Box et al. (1994) give the equation

$$\text{var}[r_k] \approx \frac{1}{N} \left(1 + 2 \sum_{k=1}^q r_k^2 \right) \quad k > q, \quad (6.2)$$

where $\text{var}[r_k]$ is the variance of the estimated autocorrelation coefficient r_k for lag k . Taking the square root of Equation (6.2) yields the large-lag standard error, which assumes that the autocorrelations r_k are nearly zero after some lag time q . The large-lag standard error provides an approximation of the standard deviation of autocorrelations r_k for lags $k > q$. In practice, if the large-lag standard error is greater than the estimated autocorrelation at lag k , then the autocorrelations beyond lag k are assumed to be small compared with this error and k is designated as the lag time of the time series. Box et al. (1994) note that in cases where the sample time series is uncorrelated i.e., when $q = 0$ and all autocorrelations at lags $k > q$ are taken to be zero, then the standard error from Equation (6.2) simplifies to $\sqrt{1/N}$.

After obtaining the decorrelation time τ , the length of the pseudorandom dataset used in the bootstrapping test, N' , is equivalent to the length of the original time series scaled by the lag time.³ Storch and Zwiers (2001) terms N' the equivalent sample size and notes that the decorrelation time is dimensionless as the time increment assumed in Equation (6.2) is generalized to be without dimension. Storch and Zwiers (2001) also warns that the implementation of such a methodology to estimate the decorrelation time, such as the large-lag standard error, is highly dependent on the problem at hand and it is unlikely that the time scales of the underlying physical process are altogether represented by the estimated decorrelation time.

³Storch and Zwiers (2001) state that for data that has a red noise spectrum, like that of the 20CRv2c daily anomalies, $N' < N$ and $\tau > 1$.

Table 6.2 gives the decorrelation time (τ) for East Hawaii and Vancouver calculated from the large-lag standard error as well as the length of the original time series (N) and the pseudorandom time series (N') to be used to compute the goodness-of-fit p-value.

Table 6.2: Locations, lengths of time series (N) and pseudorandom time series (N') and lag time (τ) of near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) 1947-2014 DJF 950 hPa air temperature anomaly time series.

Point	Location	N	N'	τ
A	East Hawaii	6137	105	59
B	Vancouver	6137	341	18

Note the large decorrelation times for both points, especially the 59 days of East Hawaii. Romanou et al. (2006) computed decorrelation times for many ocean surface flux variables using the Goddard Satellite-Based Surface Turbulent Fluxes version 2 (GSSTF2) dataset⁴ and found that, in the case of heat fluxes, the equatorial regions of the ocean basins contained the longest decorrelation time scales (approximately 90 days), suggesting the variability in the equatorial oceans is intraseasonal. Romanou et al. (2006) also finds the long decorrelation times consistent with persistent wind patterns found near the ITCZ as well as time periods related to the Madden-Julian oscillation. In the case of constructing a statistical test, the long (≈ 60 day) decorrelation time of East Hawaii results in a short pseudorandom time series to be used in the goodness-of-fit test.

Initial goodness-of-fit tests using the large-lag standard error of Equation (6.2) to estimate the decorrelation time were instructive on how a large lag time can influence the pseudorandom KS statistics used to obtain p-values. In particular, a very long decorrelation time can seriously undermine the statistical test by producing an ECDF too coarse to calculate a KS statistic relative to the observations and estimated SGS fit. This can result in a larger pseudorandom KS statistic and, when the KS statistics from all the bootstrap iterations are tabulated, a “significance bias.” Figure 6.6 illustrates pseudorandom ECDFs generated from different decorrelation times at East Hawaii and Vancouver and how a short pseudorandom time series can effect the KS statistic used to estimate the goodness-of-fit.

⁴See Chou et al. (2003) for more information on the GSSTF2 dataset.

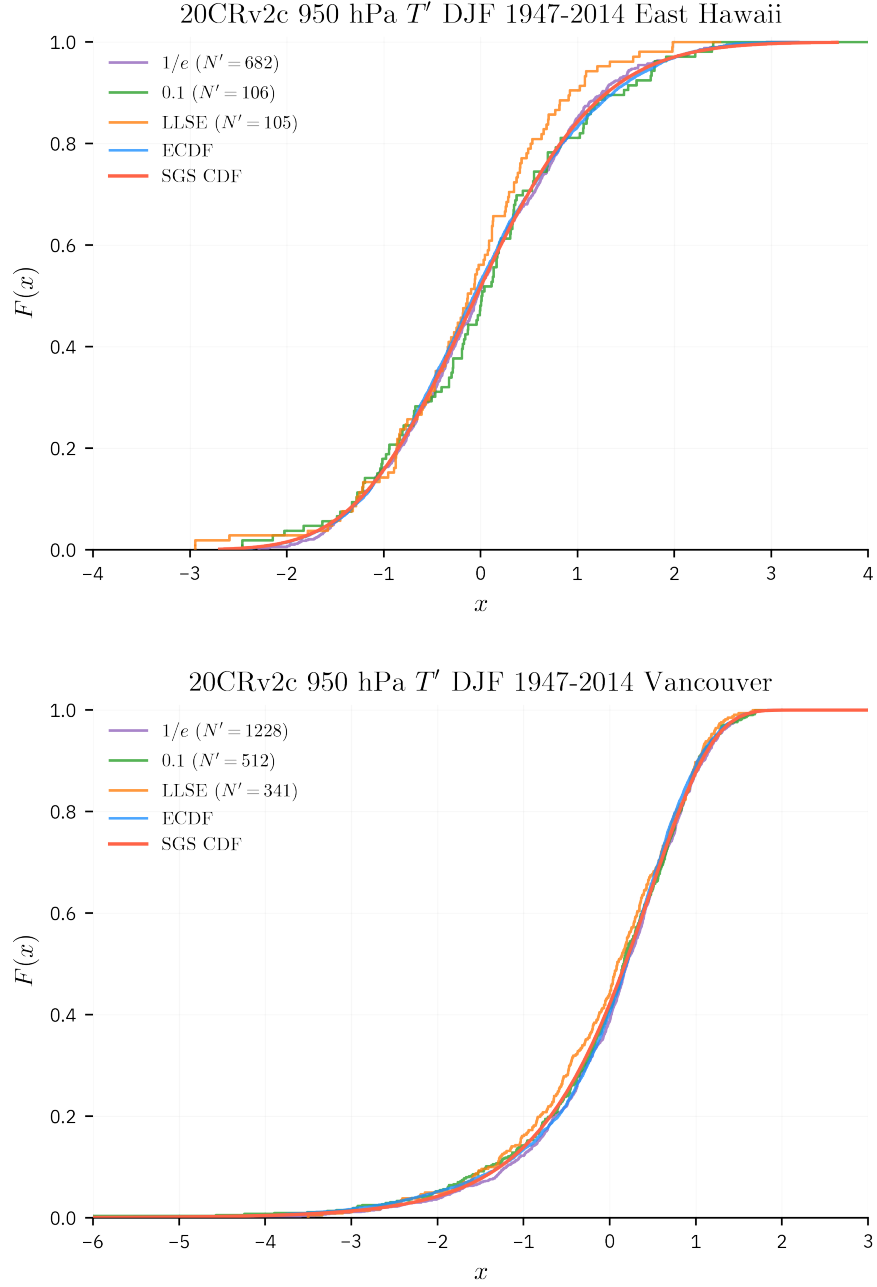


Figure 6.6: ECDF (blue) of the DJF 950 hPa 20CRv2c air temperature anomalies (1947-2014) and estimated SGS CDF (red) located at near-Gaussian East Hawaii (top, 20°N, 138°W) and Vancouver (bottom, 50°N, 124°W) compared against ECDFs obtained from pseudorandom time series scaled by decorrelation times with various cutoff thresholds ($1/e$, 0.1, large-lag standard error or LLSE). The equivalent sample size (N') for each ECDF is included.

Figure 6.6 shows the ECDFs that result from applying three different means of obtaining the decorrelation time scale. In purple is the $1/e$ autocorrelation threshold, or the lag time which the autocorrelation of the time series drops below $1/e$. The green ECDF is obtained from a pseudorandom time series whose length was determined by a 0.1 autocorrelation threshold, or the time lag at which the autocorrelation drops below 0.1. The brown ECDF is the time series scaled by the large-lag standard error, given by Equation (6.2). The equivalent sample sizes (N') for each ECDF are given in the legend. Note that as the value of N' decreases, the ECDF becomes less smooth and more “stepwise,” which would result in a larger KS statistic. In contrast, ECDFs with larger equivalent sample sizes are smoother and cluster more tightly to the estimated SGS fit. Table 6.3 gives a comparison of the decorrelation times obtained by the three methods at each point.

Table 6.3: Locations and decorrelation times computed using the $1/e$ autocorrelation threshold, 0.1 autocorrelation threshold and the large-lag standard error (LLSE) at near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) 1947-2014 DJF 950 hPa air temperature anomaly time series.

Point	Location	$1/e$	0.1	LLSE
A	East Hawaii	9	58	59
B	Vancouver	5	12	18

While it is important to appropriately calculate the decorrelation time scale so that the pseudorandom time series has an amount of information equivalent to the original data, it is more important in this study to set a high bar with which to assign a statistical measure of plausibility. Though the large-lag standard error takes care to calculate a lag time at which the autocorrelation practically vanishes, particular locations with large decorrelation times may suffer from biased goodness-of-fit tests due to extremely short pseudorandom time series. To mitigate this, the $1/e$ autocorrelation threshold will be used as the cutoff by which to calculate the decorrelation time scale. The decorrelation time scale is also limited to a maximum of 90 (days) i.e., the number of days in a DJF season. Using the $1/e$ autocorrelation threshold will generally produce shorter decorrelation time scales, which provides a longer pseudorandom time series from which to calculate the goodness-of-fit tests. In turn, this will produce a more conservative measure of statistical plausibility i.e., the KS statistic computed from the longer pseudorandom time series will be more “competitive” in the comparison

with the KS statistic obtained from the reanalysis data. Large p-values obtained using the more conservative threshold will instill further confidence in the plausibility of the SGS fit.

6.5 Results of the goodness-of-fit test

The goodness-of-fit is now calculated for the time series of near-Gaussian and non-Gaussian air temperature anomalies at points East Hawaii and Vancouver. Clauset et al. (2009) mention that if one desires the accuracy of the computed p-values to be within ϵ of the true value, then $\frac{1}{4}\epsilon^{-2}$ bootstrap iterations must be made. As such, 400 bootstrap iterations are performed which give a p-value accuracy within 0.025.

First, the KS statistic is computed between the estimated SGS distribution and time series of anomalies at each location. Then, a pseudorandom dataset drawn from the SGS distribution is obtained with a length determined from the decorrelation time obtained from calculating the lag time at the $1/e$ autocorrelation threshold. An SGS distribution is in turn fit to the pseudorandom dataset. The KS statistic of this pseudorandom SGS distribution is computed and saved. After 400 iterations, all pseudorandom KS statistics are compared with the original KS statistic obtained from the SGS fit to the respective 950 hPa air temperature anomalies and a p-value is calculated. Table 6.4 shows the results of the statistical test for both points East Hawaii and Vancouver.

Table 6.4: The KS statistic and p-value computed from the SGS goodness-of-fit test for the near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) 1947-2014 DJF 950 hPa air temperature anomaly time series.

Point	Location	KS Statistic	p-value
A	East Hawaii	0.01257	1.0
B	Vancouver	0.02497	0.5025

Even though the time series and higher moments of the near- and non-Gaussian points East Hawaii and Vancouver given in Table 6.1 are quite different, the SGS distributions fit to the time series with the procedure given in Section 5.2 are both determined to be plausible.

The bootstrapping goodness-of-fit test may be repeated using a standard normal distribution as the candidate fit in order to test how plausible the Gaussian distribution is in representing the

two time series of air temperature anomalies. The above procedure is repeated with the standard normal distribution substituted for the estimated SGS distribution, with the results displayed in Table 6.5.

Table 6.5: The KS statistic and p-value computed from the standard normal goodness-of-fit test for the near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) 1947-2014 DJF 950 hPa air temperature anomaly time series.

Point	Location	KS Statistic	p-value
A	East Hawaii	0.03	0.5425
B	Vancouver	0.1017	0.0

Employing the goodness-of-fit test for the standard normal distribution against the two locations of air temperature anomalies provides a different picture than in the SGS case. In the case of the near-Gaussian East Hawaii, the standard normal distribution is found to be plausible with a p-value of 0.5425. Inspection of the step histogram in Figure 6.2 as well as the Q-Q plot of Figure 6.3 indeed illustrates how close to Gaussian East Hawaii is. Meanwhile, the Vancouver p-value of 0 signifies that the standard normal distribution is not the correct fit to the non-Gaussian air temperature anomalies at that location.

From Section 4.1, it is stated that the standard normal distribution is itself a member of the SGS class of distributions. From Equation (4.5), in the limit case where the skewness is zero and the SGS parameter $E \rightarrow 0$, the SGS pdf converges to the standard normal distribution. Generally speaking, this means that the SGS distribution is able to represent near-Gaussian time series at least as well as the standard normal distribution. Given the p-value results of Tables 6.4 and 6.5, this seems to be the case.

While the p-values in Table 6.4 are an encouraging sign for the ability of the SGS distribution to model air temperature anomalies, Clauset et al. (2009) recommends caution in interpreting large p-values as a license to conclude that the SGS distribution is the single, correct statistical representation of the data. It is possible that other, untested distributions may yet be a better fit for the data. While the bootstrapping goodness-of-fit test can provide a measure of plausibility to an estimated SGS distribution, it cannot directly compare competing distributions against each

other. To do so requires a test statistic that performs a model selection, such as the likelihood ratio test statistic.

6.6 Model selection via likelihood ratio test

The likelihood ratio test is a statistical test that, as the name implies, uses a ratio of the likelihoods of two competing distributions to determine the relative goodness-of-fit. This test assumes that one of the candidate distributions is nested in the other, meaning that one model can be constructed by imposing constraints on the other. Here, the fitted SGS distribution is compared to the standard normal distribution as a test of non-Gaussianity. As mentioned previously, the standard normal distribution is a member of the SGS class of distributions. This means it is nested within the SGS distribution, as from Equation (4.5), in the limit case where the skewness is zero and the SGS parameter $E \rightarrow 0$, the SGS pdf converges to the standard normal distribution.

The likelihood ratio test statistic, LR , may be written in terms of the SGS and Gaussian distributions as

$$LR = 2 \ln \left[\frac{\mathcal{L}_{\text{SGS}}}{\mathcal{L}_{\text{Gauss}}} \right] = 2(\ell_{\text{SGS}} - \ell_{\text{Gauss}}), \quad (6.3)$$

where \mathcal{L} is the likelihood function and ℓ is the log-likelihood function of the respective distributions. Here, the SGS distribution is the weakly constrained (more free parameters) distribution and the standard normal distribution is the heavily constrained (no free parameters) model. Both the SGS and standard normal pdfs, given in Equation (6.4) respectively as $p(x)$, are used in specifying log-likelihood functions, where

$$\begin{aligned} \ell(X)_{\text{Gauss}} &= \sum_{i=1}^N p(x_i), \\ \ell(E, b, g | X)_{\text{SGS}} &= \sum_{i=1}^N p(x_i, E, b, g), \end{aligned} \quad (6.4)$$

where N is the length of the time series x_1, x_2, \dots, x_N and the SGS parameters E , b and g are estimated from the time series. The nested condition of the likelihood ratio test means that the SGS distribution is able to represent near-Gaussian time series at least as well as the standard normal distribution. However, the likelihood ratio test aims to determine if the difference in the fit of the two models is significant and unlikely to have resulted from statistical fluctuations (Wilks

(2006)). If difference is shown to be significant, then the weakly constrained model i.e., the SGS distribution, is said to be the most plausible fit.

According to Wilks (2006), given a large enough sample size, the sampling distribution of LR in Equation (6.3) is χ^2 distributed. The degrees of freedom of the χ^2 distribution is determined from the difference in the estimated parameters. In the case of the SGS distribution and the standard normal distribution, this difference is 3. Conveniently, a p-value for the likelihood ratio test may be found by computing the pdf of the χ^2 distribution with three degrees of freedom at the location of the LR statistic. The pdf of the χ^2 distribution with three degrees of freedom may be written as

$$p(x) = \begin{cases} \frac{x^{\frac{1}{2}} e^{-\frac{x}{2}}}{2^{\frac{3}{2}} \Gamma(\frac{3}{2})}, & x > 0; \\ 0, & \text{elsewhere.} \end{cases} \quad (6.5)$$

Here, $\Gamma(\frac{3}{2})$ is the gamma function (Walpole et al. (2017)). In the SGS case, if the fit of the SGS distribution is favorable and the difference between the SGS fit and the standard normal distribution is not merely due to statistical fluctuations, then the p-value obtained from the χ^2 pdf of Equation (6.5) will be close to zero. In instances where the SGS distribution is strongly preferred, the p-value may be extremely small and approximately zero. In contrast to the p-value obtained from the goodness-of-fit test, a large p-value (again at least greater than 0.1) approaching 1 is indicative of a sufficiently improbable SGS model preference. Note that in the case of where the Gaussian distribution is preferred to the SGS distribution, LR will be negative and the pdf of the χ^2 distribution will be 0. To distinguish between p-values of the negative LR case and the large positive LR case, the p-value in the negative LR case is assigned a value of 1.

Applying the likelihood ratio test to compare the SGS distribution fit at points East Hawaii and Vancouver to the standard normal distribution yields the results given in Table 6.6.

Table 6.6: Likelihood ratio test statistic of the fitted SGS distribution against the standard normal distribution for the near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) 1947-2014 DJF 950 hPa air temperature anomaly time series.

Point	Location	LR	p-value
A	East Hawaii	41.82	2.52×10^{-18}
B	Vancouver	828.62	≈ 0

Note that the magnitude of LR is non-dimensional and does not have a particular significance, other than the sign of the LR statistic signifying a preference of the data for either the SGS (positive) or standard normal (negative) distribution. A LR magnitude much larger than zero represents a more plausible fit, but the computation of the p-value provides the standardized metric of comparison.

Taking into account the visual inspection of the Q-Q plots, the goodness-of-fit test and the model comparison of the likelihood ratio test, it appears that the SGS distribution is a plausible fit for the 950 hPa air temperature anomalies at both the near-Gaussian East Hawaii and the non-Gaussian Vancouver. In the next chapter, the statistical analysis of the SGS distribution is expanded to more 20CRv2c variables and a global spatial domain.

CHAPTER 7

GLOBAL SGS EVALUATIONS OF 20CRV2C VARIABLES

After outlining the statistical methodology for evaluating the fit of an estimated SGS distribution in Chapter 6, the techniques are applied to multiple 20CRv2c variables and evaluate the plausibility of the SGS distribution on a global scale. For reference, all variables in this chapter are persistent wintertime (DJF) anomalies taken from the 20CRv2c dataset from 1947-2014, standardized according to the process described in Section 2.1.

For each variable, an SGS distribution is fit to the time series of anomalies at each grid point, using the method of moments described in Section 4.2 and implemented in Section 5.3. An assessment of the SGS model fit at each grid point is also provided, where violations of the constraints governing the SGS distribution during the distribution fit are tabulated. The constraint violations of the SGS distribution depicted here follow those of Section 4.2, where the model may be violated in the following ways:

1. $E^2 < 0$: E and g are both negative, violating the ϵ -constraint of Equation (4.10)
2. $E > E_{\max}$, or the kurtosis does not exist.
3. $b^2 < 0$: E is positive but b is negative, violating the ϵ -constraint of Equation (4.10)
4. $|\text{skew}| > \text{skew}_{\max}$: the magnitude of the sample skewness is greater than $\sqrt{32}$, the maximum allowable under the constraints on E

The constraint violations listed are each handled in accordance with the methodology described in Section 4.2.

The decorrelation time scale is also computed using the $1/e$ autocorrelation threshold described in Section 6.4 to provide the equivalent sample size N' required for the pseudorandom time series used in the bootstrapping goodness-of-fit test. The decorrelation time scale and SGS distribution of each grid point are then used to compute the goodness-of-fit test for the estimated SGS distribution relative to the data, following the methodology outlined in Section 6.3. After bootstrapping 100 pseudorandom datasets and tabulating each KS statistic relative to the estimated SGS distribution,

p-values at each grid point may be found. The goodness-of-fit test is also repeated using the Gaussian distribution, and the p-value results from both goodness-of-fit tests are displayed as global maps. Points that are shaded blue correspond to an implausible SGS fit, where the p-value is less than the arbitrary cutoff of 0.1. Admittedly, according to the Clauset et al. (2009) approximation of the p-value accuracy ϵ discussed in Section 6.5, 100 bootstrap iterations only provide a p-value accuracy within 0.05. However, tests with increased numbers of bootstrap iterations become computationally expensive and experiments have shown that the overall spatial p-value structure is unchanged when using a larger amount of pseudorandom datasets. Nonetheless, using a larger amount of bootstrap iterations will provide greater test confidence.

Next, a comparison of the values of skewness and kurtosis computed at each grid point is made relative to the goodness-of-fit p-value obtained through the bootstrapping significance test. Again, points that are shaded blue correspond to an implausible SGS fit, where the p-value was calculated to fall below the 0.1 significance threshold. Finally, the likelihood ratio test statistic described in Section 6.6 is used to perform a model selection test, computed at each grid point, which determines if the estimated SGS distribution is a statistically preferred fit to the data, relative to the Gaussian distribution. Points that are shaded blue again correspond to an implausible SGS fit, where the SGS distribution is either not preferred or statistically distinguished from the standard normal distribution.

Note that for both the goodness-of-fit test and likelihood ratio test statistic, p-value threshold values near 0.1 are shaded in incrementally lighter shades of blue to convey a sliding scale of plausibility rather than a particular threshold of binary significance. For reference, Table 7.1 provides a simplified guide to the p-value plausibility thresholds and map colors used in the following analysis.

Table 7.1: Lookup table for the p-value significance thresholds, plausible (“good” fit) and implausible (“bad” fit) values and map colors for the goodness-of-fit (GOF) and the SGS-to-Gaussian comparison of the likelihood ratio test statistic (LRTS) statistical tests.

Test	p-value threshold	“good” values	“good” color	“bad” values	“bad” color
GOF	0.1	$p > 0.1$	gray	$p < 0.1$	blue
LRTS	0.1	$p < 0.1$	gray	$p > 0.1$	blue

7.1 950 hPa air temperature

Beginning with the 1947-2014 DJF 950 hPa air temperature anomalies that were investigated in Chapter 6, fitting an SGS distribution at each grid point yields the parameters E , b and g which are shown in Figure 7.1.

The parameter maps of Figure 7.1 are complex, but as the method of moments requires the variance, skewness and kurtosis to compute the SGS parameters (see Equation (4.8) and Section 5.2), it is instructive to compare them to the maps of skewness and kurtosis given in Figure 6.1. Parameter E has a larger overall magnitude in regions of greater positive kurtosis, but there is also a tendency toward slightly higher values when kurtosis is near zero while the magnitude of the skewness is increased. Parameter b meanwhile has large magnitudes where either the value of E is small, the skewness is small, or a combination of both. Finally, parameter g is the only parameter that is not positive, meaning that the sign of g determines the sign of the skewness. As such, negative values of g correspond to negative skewness while positive values of g correspond to positive skewness. Near-zero values of g typically indicate regions of near-Gaussianity.

Figure 7.2 depicts the types of SGS model violations that occurred during the distribution fit, while Figure 7.3 shows the decorrelation time scales computed at all 950 hPa air temperature anomaly grid points. The model violations in Figure 7.2 are all generally of the $E^2 < 0$ type, save for a few violations where parameter g was not positive. A comparison of the violations with the air temperature skewness and kurtosis maps of Figure 6.1 indicates that the violations are typically limited to regions of negative kurtosis (which the SGS distribution cannot fit without violating constraints).

From Figure 7.3, it appears that while relatively short decorrelation times of approximately one week are found in the subtropics, mid-latitudes and polar regions, the tropics have a monthly to seasonal lag time. Caution is required, however, with the large lag times in the tropics. While perhaps representative of the persistence statistics found near the ITCZ and involving the Madden-Julian oscillation, large lag times may bias p-value statistics in a “distribution-plausible” way (see discussion in Section 6.4).

Both the estimated SGS fit and the decorrelation time scale are used to produce a goodness-of-fit test that yields p-values at each grid point. The goodness-of-fit test comparing both the fit of the

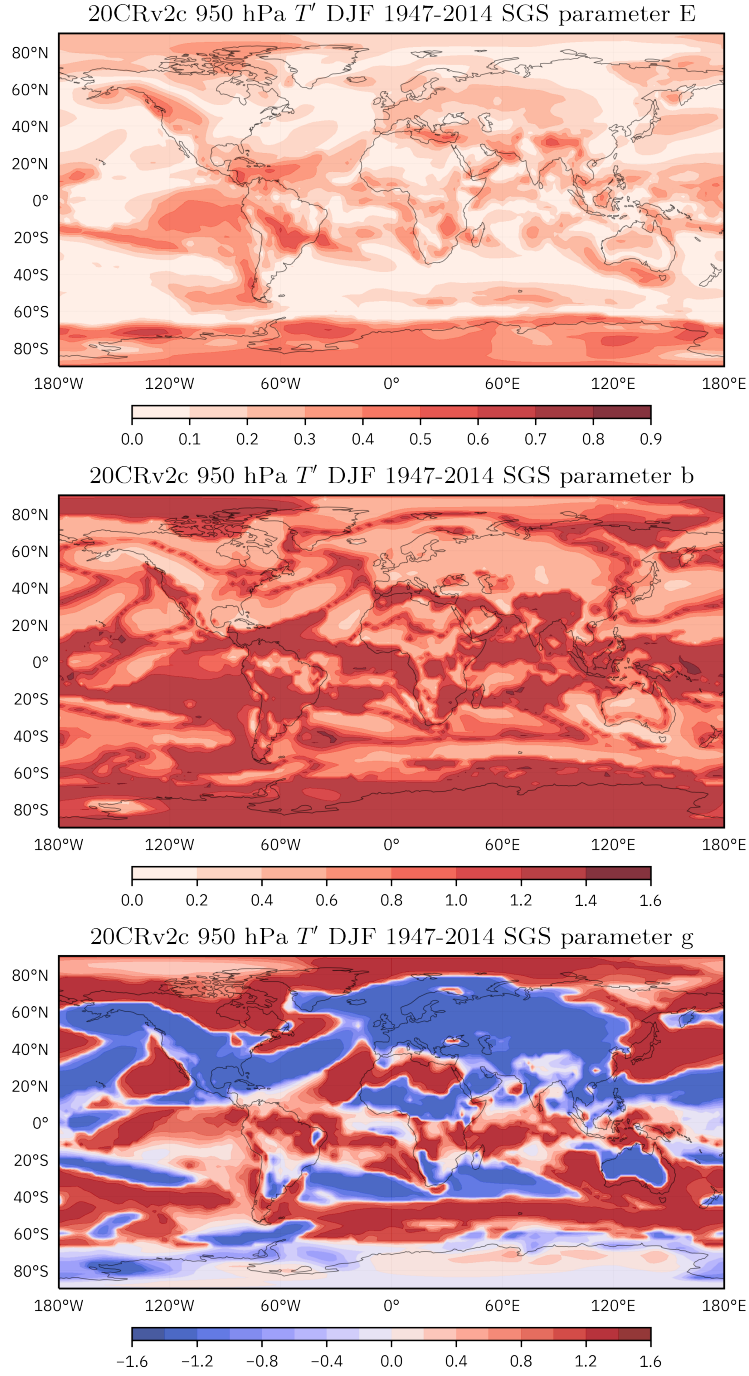


Figure 7.1: Maps of SGS parameters E (top), b (middle) and g (bottom) estimated from standardized DJF 950 hPa 20CRv2c air temperature anomalies, 1947-2014.

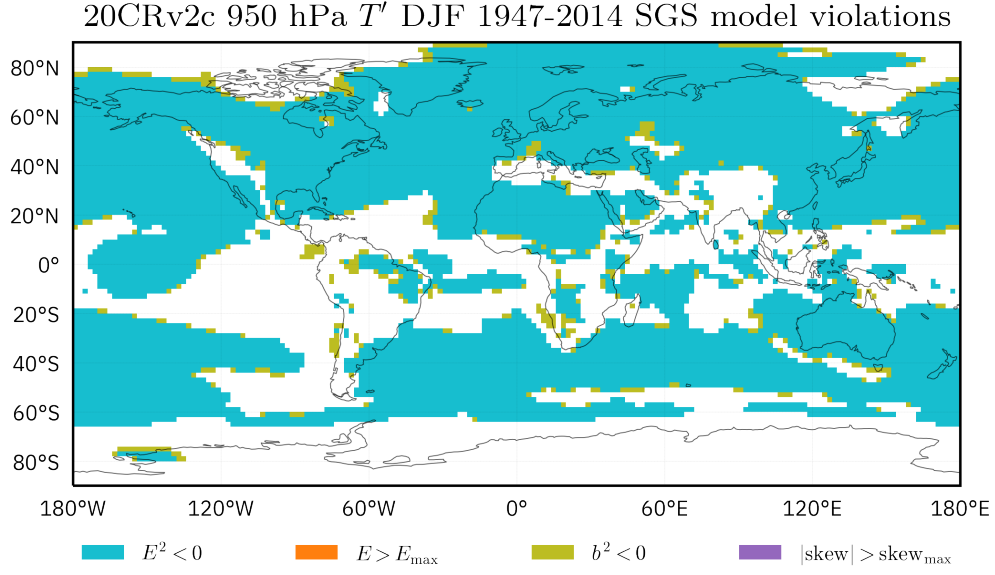


Figure 7.2: Map of the SGS model violations encountered when fitting an SGS distribution to time series of standardized DJF 950 hPa 20CRv2c air temperature anomalies, 1947-2014. Violations $E^2 < 0$ (cyan) and $b^2 < 0$ (olive) are most prominent.

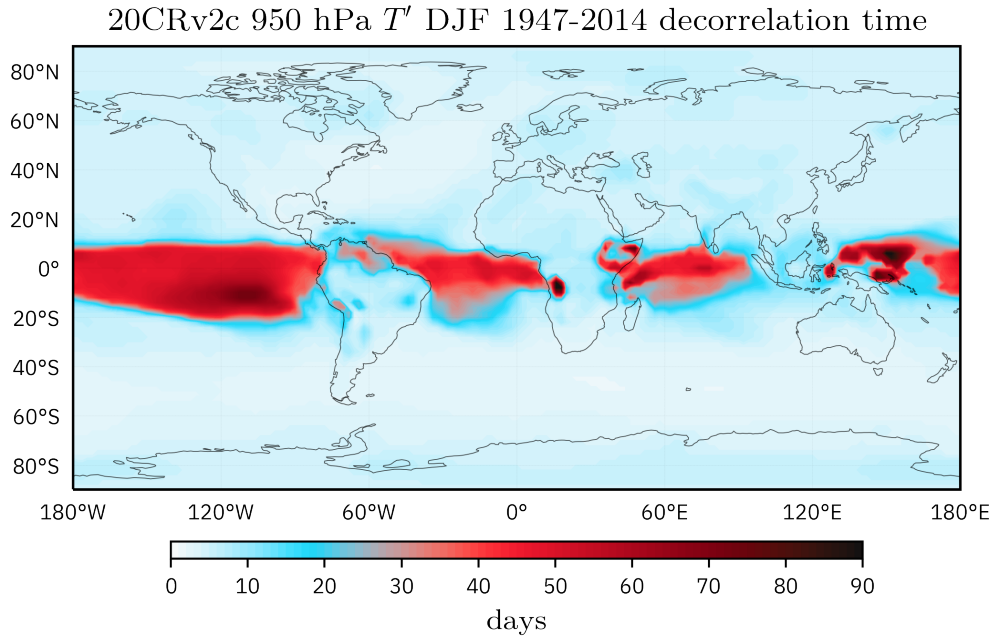


Figure 7.3: Decorrelation time scale in days computed using the $1/e$ autocorrelation threshold for the standardized DJF 950 hPa 20CRv2c air temperature anomalies, 1947-2014.

SGS distribution and the standard normal distribution to the air temperature anomalies are shown in Figure 7.4.

Note that the grid points colored blue in Figure 7.4 correspond to the estimated SGS or standard normal distribution (respectively) being a implausible fit to the data. It is important to compare the results of the goodness-of-fit tests given in Figure 7.4 to the maps of skewness and kurtosis given in Figure 6.1. Upon inspection, it appears that regions where the estimated SGS distribution is ruled to be implausible generally have negative kurtosis, or are in near-zero kurtosis regions with a non-zero magnitude of skewness. A poorer SGS fit in regions of negative kurtosis is expected, as the skewness-kurtosis inequality given in Equation (4.9) is violated in cases of time series with a negative sample kurtosis, meaning that the estimation of parameter E via method of moments will not be optimal. In contrast, the standard normal distribution typically has an implausible fit to the data in regions where the skewness and kurtosis are large. However, the implausible region of the standard normal distribution does not seem to include the tropics, where large regions of positive kurtosis exist.

The plausibility of the standard normal distribution in the non-Gaussian tropics is likely due to the large decorrelation times scales of the 950 hPa air temperature anomalies in the tropical ocean regions. For example, Figure 7.5 shows the bootstrap CDFs generated (in olive green) via pseudorandom time series during the goodness-of-fit test which compares the standard normal distribution to the data. The KS statistic computed from the bootstrap CDFs here is likely to be inferior to that of the standard normal curve (as discussed in Section 6.4), as the decorrelation time scale of 55 days translates to a pseudorandom time series length of 112. Even though the non-Gaussianity found in the tropical oceans should favor the plausibility of the estimated SGS distribution, the p-values obtained through the goodness-of-fit test will likely be larger due to the long decorrelation times. As such, caution should be taken for the goodness-of-fit results located in the tropical ocean basins.

Figure 7.6 compares the values of skewness and kurtosis computed at each grid point relative to the goodness-of-fit p-value obtained through the bootstrapping significance test. Points that are shaded blue correspond to an implausible estimated SGS fit. The solid black parabolic line represents the skewness-kurtosis constraint of Equation (4.9), where the solid gray line is the same constraint, only with a vertical offset applied to align with the data. The dashed line corresponds

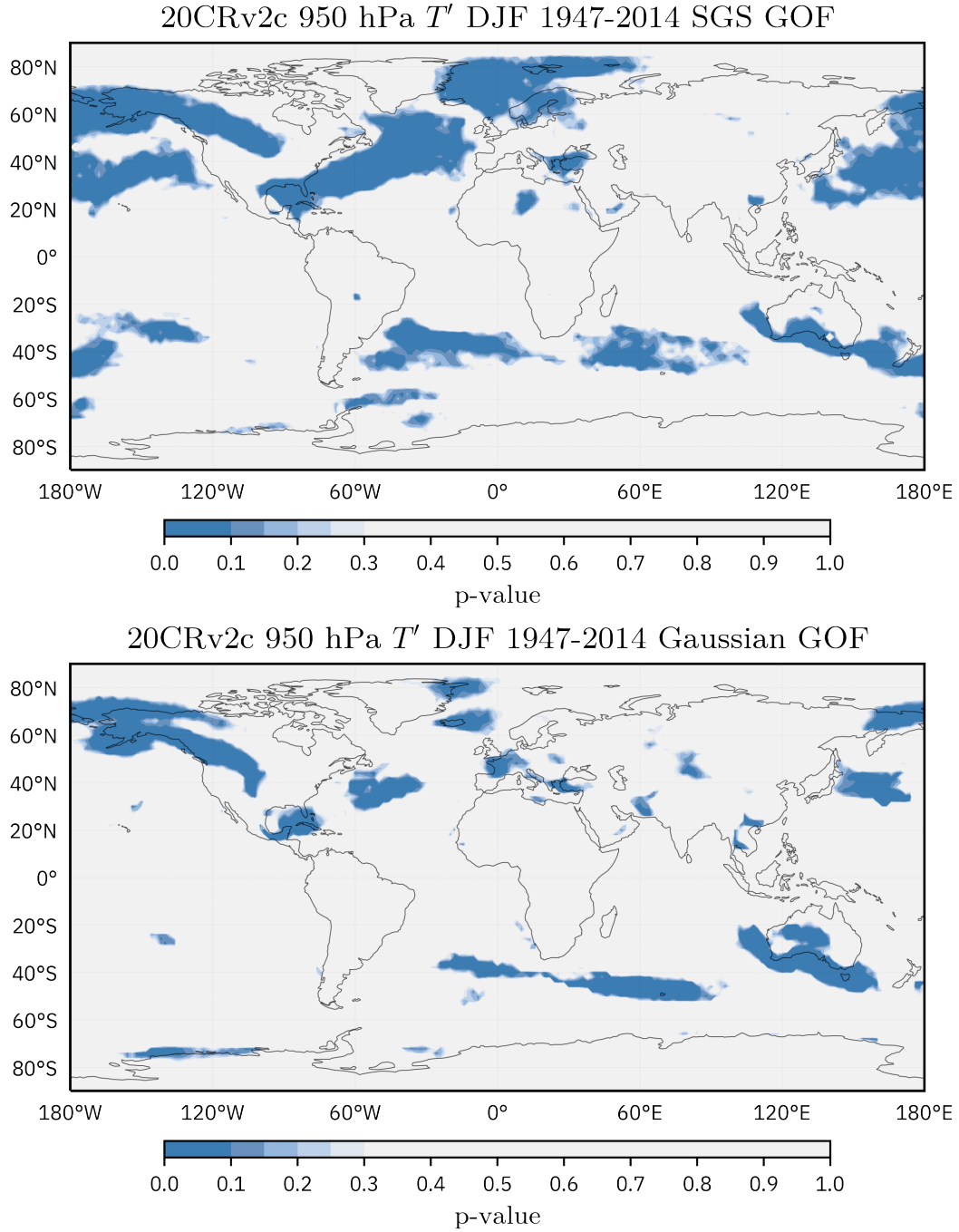


Figure 7.4: p-values for the estimated SGS (top) and standard normal (bottom) goodness-of-fit (GOF) test computed using 100 bootstrap iterations for all gridded time series of standardized DJF 950 hPa 20CRv2c air temperature anomalies, 1947-2014. Gray points correspond to a plausible SGS fit, while points gridded blue, especially those with values less than 0.1, correspond to an implausible SGS fit.

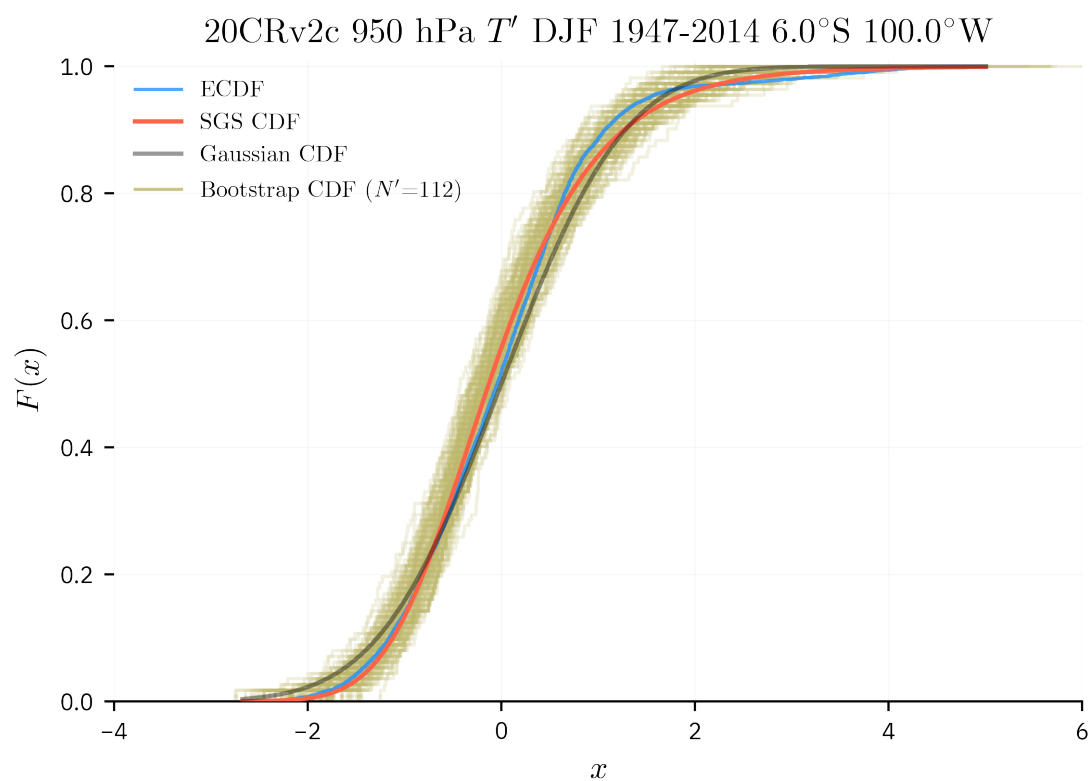


Figure 7.5: Bootstrap CDFs (olive) generated during the standard normal goodness-of-fit (GOF) test for an equatorial time series of 950 hPa 20CRv2c air temperature anomalies, 1947-2014, with an equivalent sample size (N') of 112. The estimated SGS CDF (red) is compared with the ECDF of the time series of anomalies (blue). The standard normal CDF (gray) is provided for reference.

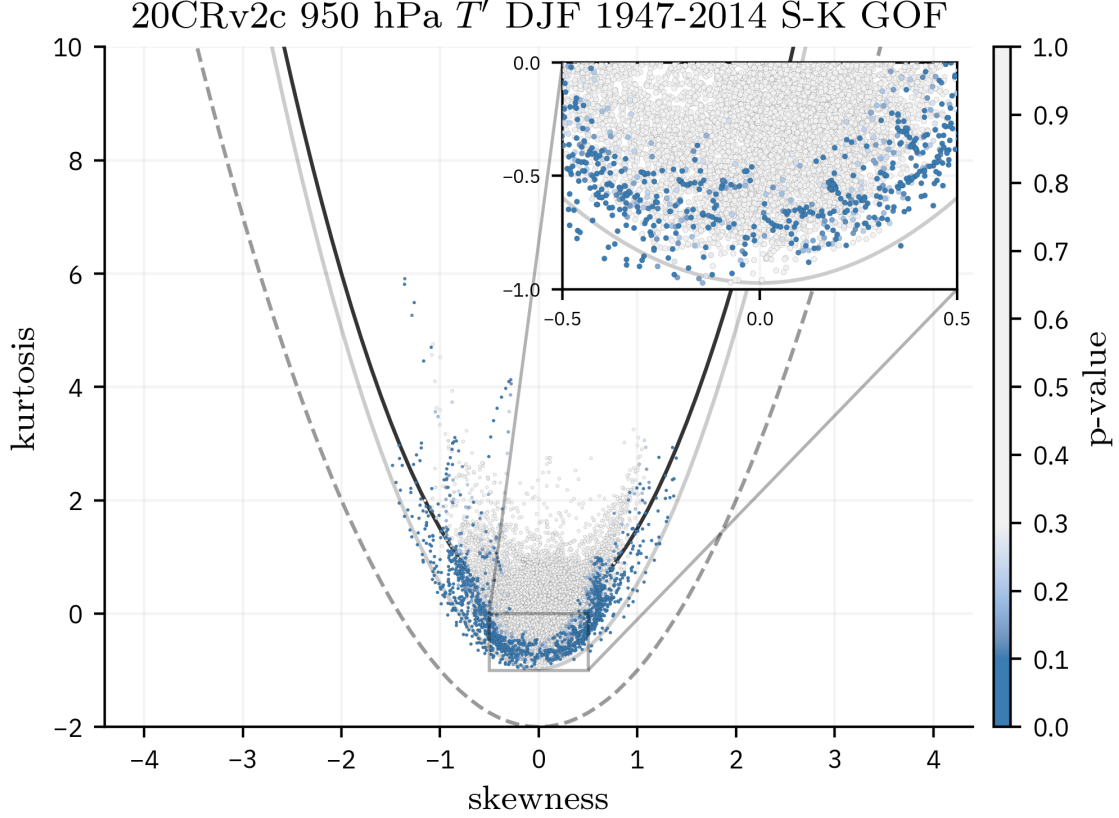


Figure 7.6: Skewness-kurtosis scatter plot showing the goodness-of-fit (GOF) p-value relative to the sample skewness and kurtosis computed for each time series of 950 hPa 20CRv2c air temperature anomalies, 1947-2014. The solid black parabola indicates the skewness-kurtosis relationship $kurt \geq \left(\frac{3}{2}\right) skew^2$, where the light gray solid parabola is $kurt \geq \left(\frac{3}{2}\right) skew^2 - r$ where r is empirically fit to the data. The dotted gray line represents the $kurt = skew^2 - 2$ constraint.

to the general constraint imposed on the skewness-kurtosis relationship, where $K \geq S^2 - 2$ (Wilkins (1944), Sardeshmukh and Sura (2009)). It is plain from Figure 7.6 that the overwhelming quantity of points that are judged to be an implausible fit to the time series of air temperature anomalies violate the skewness-kurtosis constraint of Equation (4.9). In these cases, parameter E is modified to bring the estimated SGS distribution back into agreement with the skewness-kurtosis constraint at the cost of the fit’s accuracy. Generally speaking, points with values of skewness and kurtosis that fall outside of the black parabola of Figure 7.6 are typically represented by implausible SGS distributions, while plausible SGS distributions are typically found within the black parabola.

Neglecting the issue of small bootstrap time series for grid points in the tropical oceans, Figure

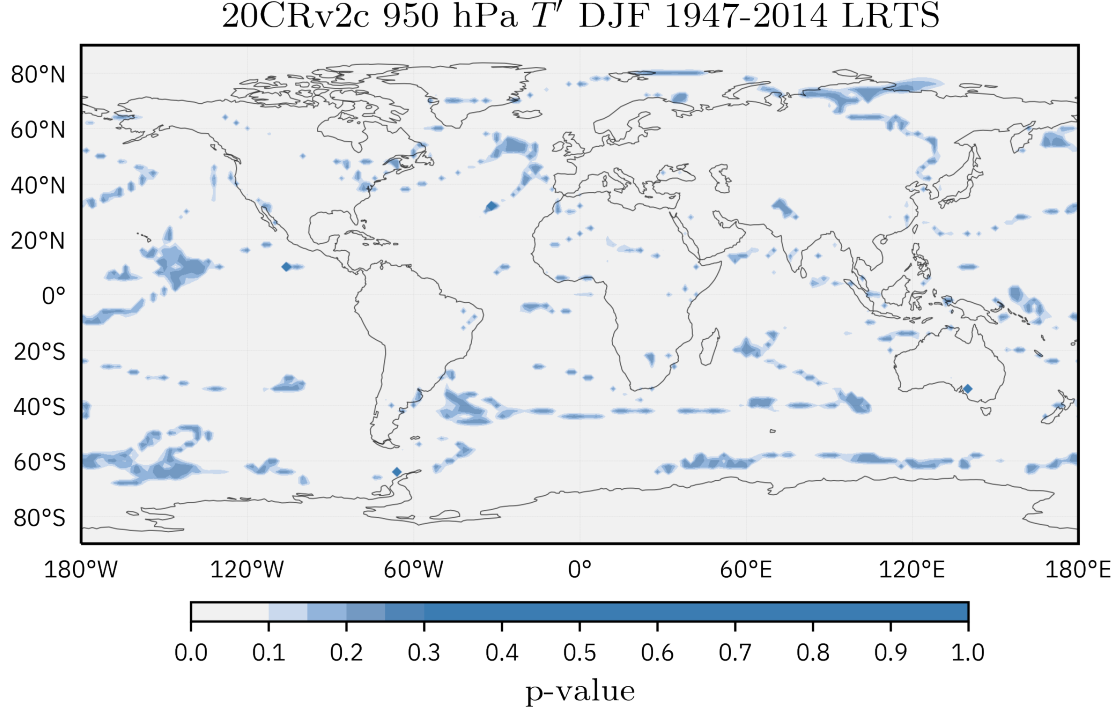


Figure 7.7: p-values of the likelihood ratio test statistic (LRTS) comparison between the estimated SGS and standard normal distributions for all gridded time series of standardized DJF 950 hPa 20CRv2c air temperature anomalies, 1947-2014. Gray points correspond to grid points where the SGS distribution is statistically preferred to the standard normal distribution. Points gridded blue, especially those with values much greater than 0.1, correspond to a statistically insignificant preference for the SGS fit or a preference for the standard normal distribution.

7.4 shows that the standard normal distribution in total has a greater percentage of plausible grid points where the p-value is greater than 0.1. However, this does not mean that the standard normal distribution is superior in fit to the SGS distribution. To address the question of which distribution is preferred by the data, a likelihood ratio test is computed according to Section 6.6 comparing the likelihood of the estimated SGS distribution to the likelihood of the standard normal distribution, with the results shown in Figure 7.7.

The p-values obtained from the likelihood ratio test in Figure 7.7 show a statistical preference for the estimated SGS distribution at a grid point if the p-value is close to zero. To simplify analysis, a threshold of 0.1 is used to distinguish between a plausible and implausible fit, where values greater than 0.1 (especially much greater) are reflective of an SGS distribution that does not

fit the data as well as the standard normal distribution. From the map, it is quite apparent that the estimated SGS distribution far outperforms the standard normal distribution in statistically representing the time series of air temperature anomalies at each grid point. The regions where the SGS distribution is not preferred to the standard normal distribution are limited to regions that share both a near zero skewness and kurtosis that is equal to or less than zero. While this is a new result, this is also theoretically expected as the SGS distribution has more degrees of freedom than the standard normal distribution, where the SGS distribution is able to converge to the standard normal distribution as $E \rightarrow 0$. It is therefore no surprise that the SGS distribution can represent near-Gaussian distributions just as well as the standard normal distribution, provided the sample time series does not have a skewness and a zero or negative kurtosis. Finally, as the likelihood ratio test statistic is a direct comparison of statistical models using the log-likelihoods of the two distributions, no bootstrapping is needed and the larger decorrelation time scales seen in the tropical ocean basins are not a cause of statistical concern.

7.2 500 hPa geopotential height

As in the 950 hPa air temperature anomalies, the SGS distribution is fit to each time series of 500 hPa geopotential height anomalies, producing an SGS distribution at each grid point (not shown here). A map of the model violations encountered in fitting the geopotential height anomalies is given in Figure 7.9. Similar to the violations encountered in the fitting of the 950 hPa air temperature anomalies, the primary regions of difficulty correspond to areas of negative kurtosis, seen in Figure 7.8.

The decorrelation time scales for all grid points are then computed producing Figure 7.10. Like the decorrelation time scale field of the 950 hPa air temperature anomalies in Figure 7.3, long decorrelation time scales exist in the tropics. However, unlike those of the air temperature anomalies, the large decorrelation times are found over all tropical longitudes, not just those over ocean basins. These large decorrelation times serve as a warning on interpreting later goodness-of-fit results in the tropics in an overly favorable light. The p-value results of the goodness-of-fit test for the estimated SGS distributions and the Gaussian distribution are shown in Figure 7.11.

From Figure 7.11 and inspection of the 500 hPa geopotential height skewness and kurtosis maps of Figure 7.8, the relationship of large SGS goodness-of-fit p-values (and small standard normal

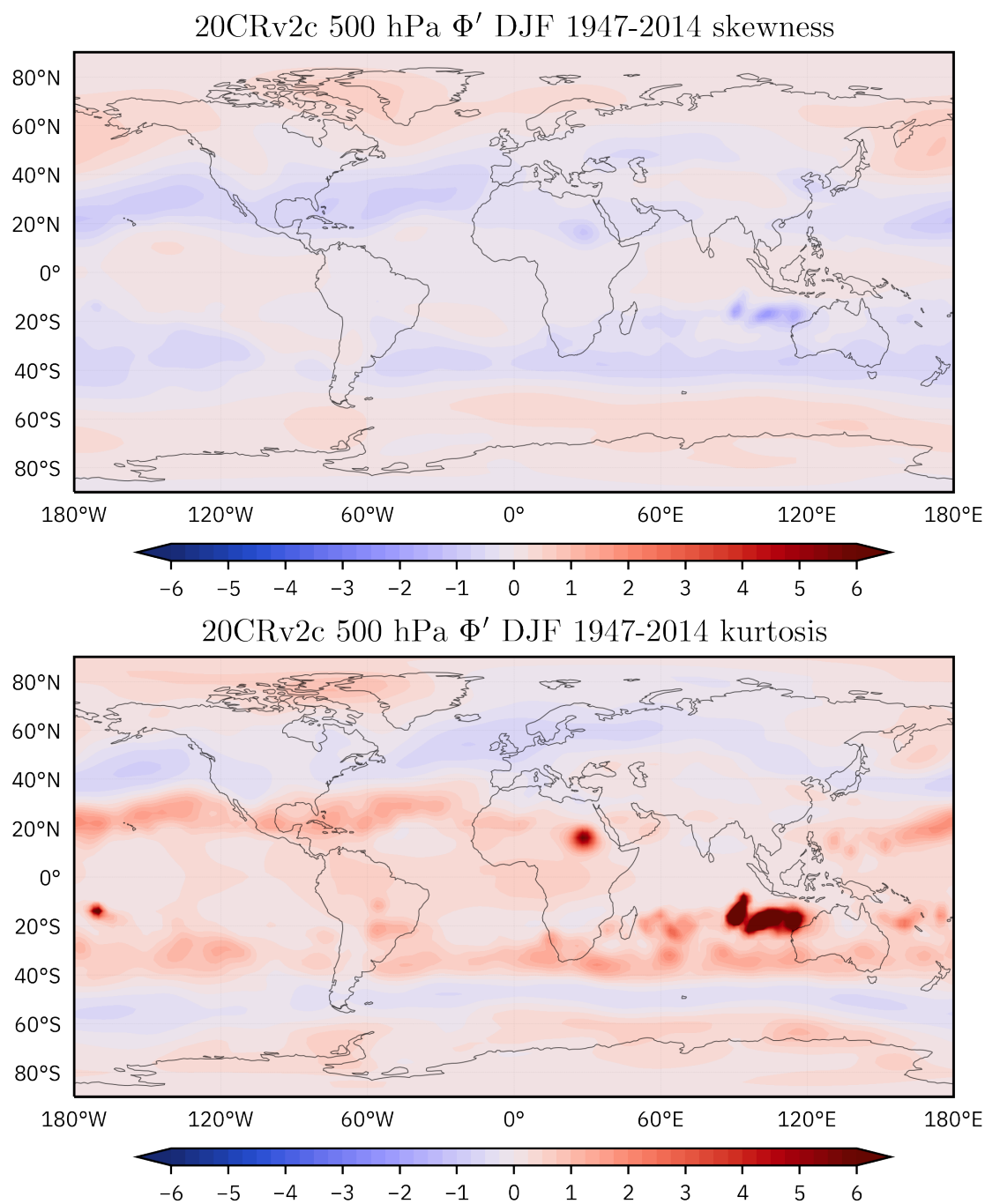


Figure 7.8: Maps of skewness and kurtosis of standardized DJF 500 hPa geopotential height anomalies from the 20CRv2c dataset, 1947-2014.

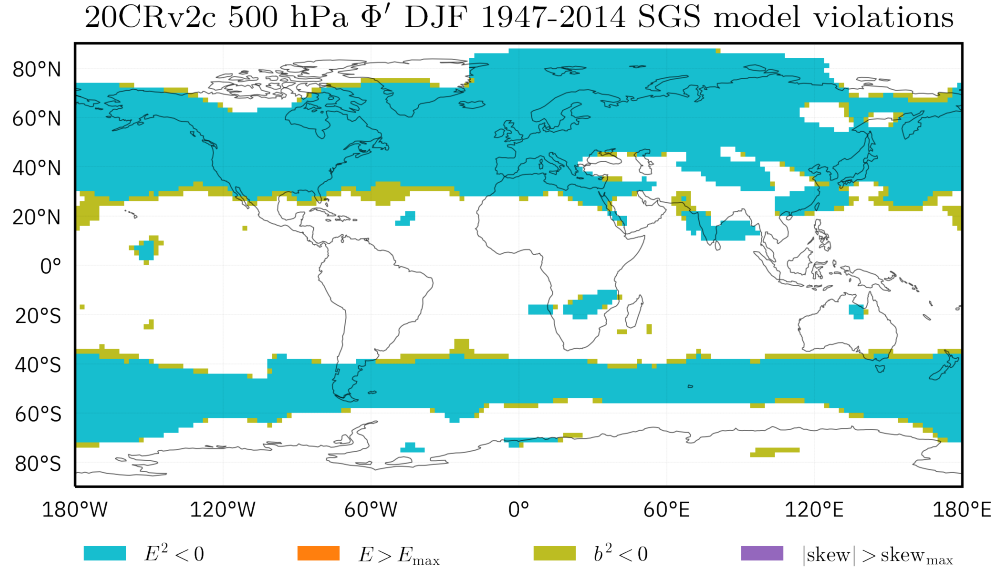


Figure 7.9: Map of the SGS model violations encountered when fitting an SGS distribution to time series of standardized DJF 500 hPa 20CRv2c geopotential height anomalies, 1947-2014. Violations $E^2 < 0$ (cyan) and $b^2 < 0$ (olive) are most prominent.

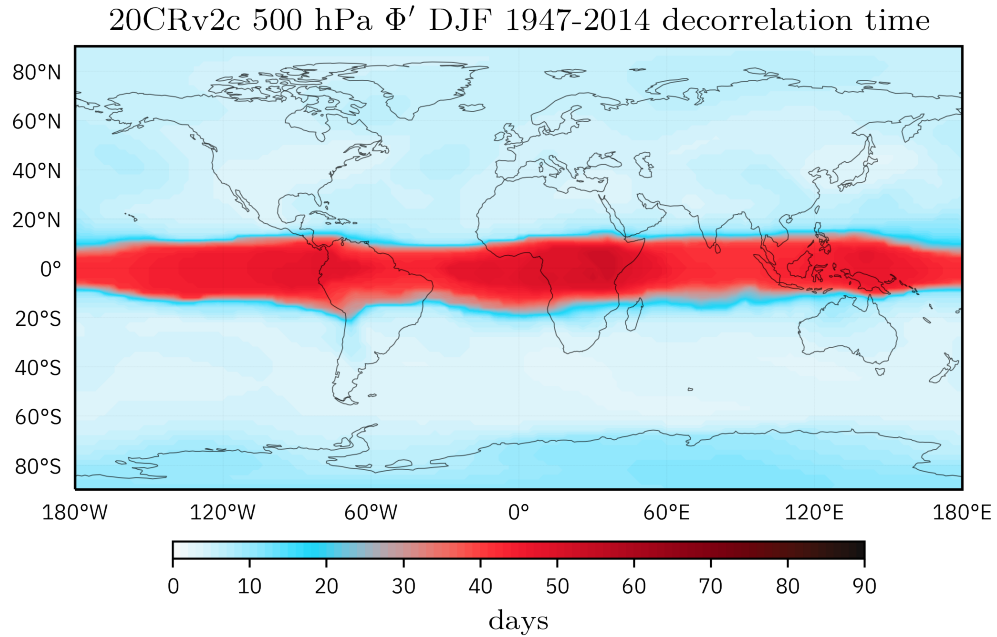


Figure 7.10: Decorrelation time scale in days computed using the $1/e$ autocorrelation threshold for the standardized DJF 500 hPa 20CRv2c geopotential height anomalies, 1947-2014.

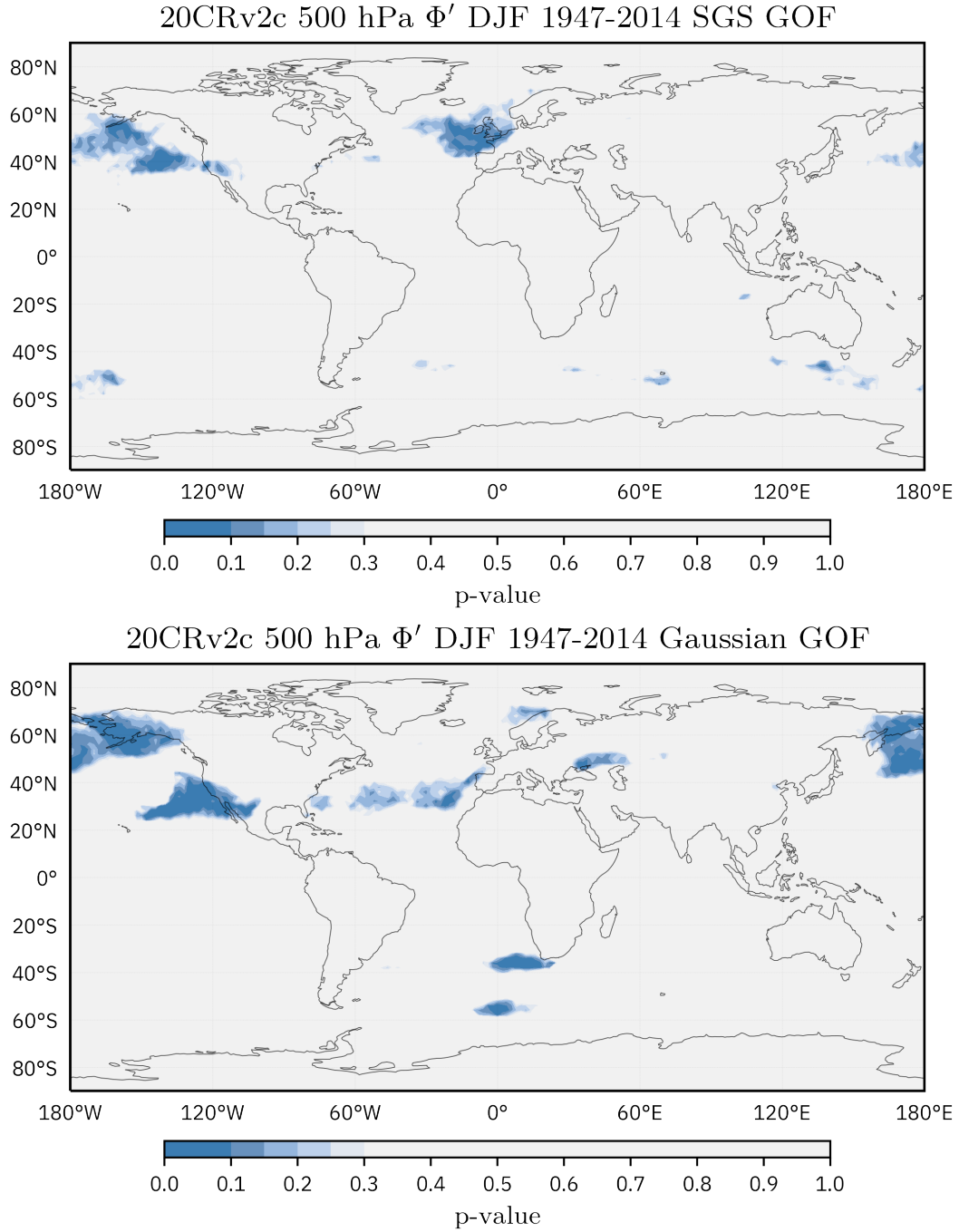


Figure 7.11: p-values for the estimated SGS (top) and standard normal (bottom) goodness-of-fit (GOF) test computed using 100 bootstrap iterations for all gridded time series of standardized DJF 500 hPa 20CRv2c geopotential height anomalies, 1947-2014. Gray points correspond to a plausible SGS fit, while points gridded blue, especially those with values less than 0.1, correspond to an implausible SGS fit.

p-values) with larger magnitudes of higher moments is observed. Conversely, implausible fits of the data to the SGS distribution are found in regions of negative kurtosis, where the fitting of SGS parameters via the method of moments violates the constraints of the SGS distribution. Note that in Figure 7.11 there are no implausible fits in the tropics. This might be expected in the SGS case, as the kurtosis is nearly entirely positive there. However, no tropical regions of implausibility in standard normal case point to a decorrelation time scale bias. Outside the tropics, the plausibility of the estimated SGS fits (as well as those of the standard normal distribution) are widespread.

Figure 7.12 describes the relationship of skewness and kurtosis relative to the plausibility of the SGS fit. Again, the time series that are not plausibly represented by an SGS distribution tend to be in regions of negative kurtosis, or where the skewness-kurtosis constraint of Equation (4.9) is strongly violated.

Finally, a likelihood ratio test is computed comparing the estimated SGS distribution to that of the standard normal distribution, with the results shown in Figure 7.13. As in the 950 hPa air temperature likelihood ratio test, only in regions of near-Gaussianity does the estimated SGS distribution not produce a statistically plausible fit when compared to the standard normal distribution, save a small group of large p-values associated with a pocket of very large kurtosis in the southeast Indian Ocean. Interestingly, one can pick out the mean jet stream location in both the northern and southern hemispheres by the mid-latitudinal band of large p-values. The mean storm track is represented in the skewness field, where the poleward positive skewness values adjacent to the equatorward negative skewness values produce a zero-skewness contour.

7.3 300 hPa relative vorticity

Figure 7.14 provides maps of skewness and kurtosis for the 300 hPa DJF relative vorticity anomalies from 1947-2014. The model violations from the fitting of SGS distributions to the gridded time series of 300 hPa relative vorticity anomalies are provided in Figure 7.15. With the same types of model violations as the previous variables, it is evident that the SGS distribution cannot be fit to a time series with negative sample kurtosis without violating, with some magnitude, the constraints imposed on the SGS distribution itself.

Figure 7.16 shows the decorrelation time scales for the field of 300 hPa relative vorticity anomalies. Unlike those of the 950 hPa air temperature or 500 hPa geopotential height anomalies, the

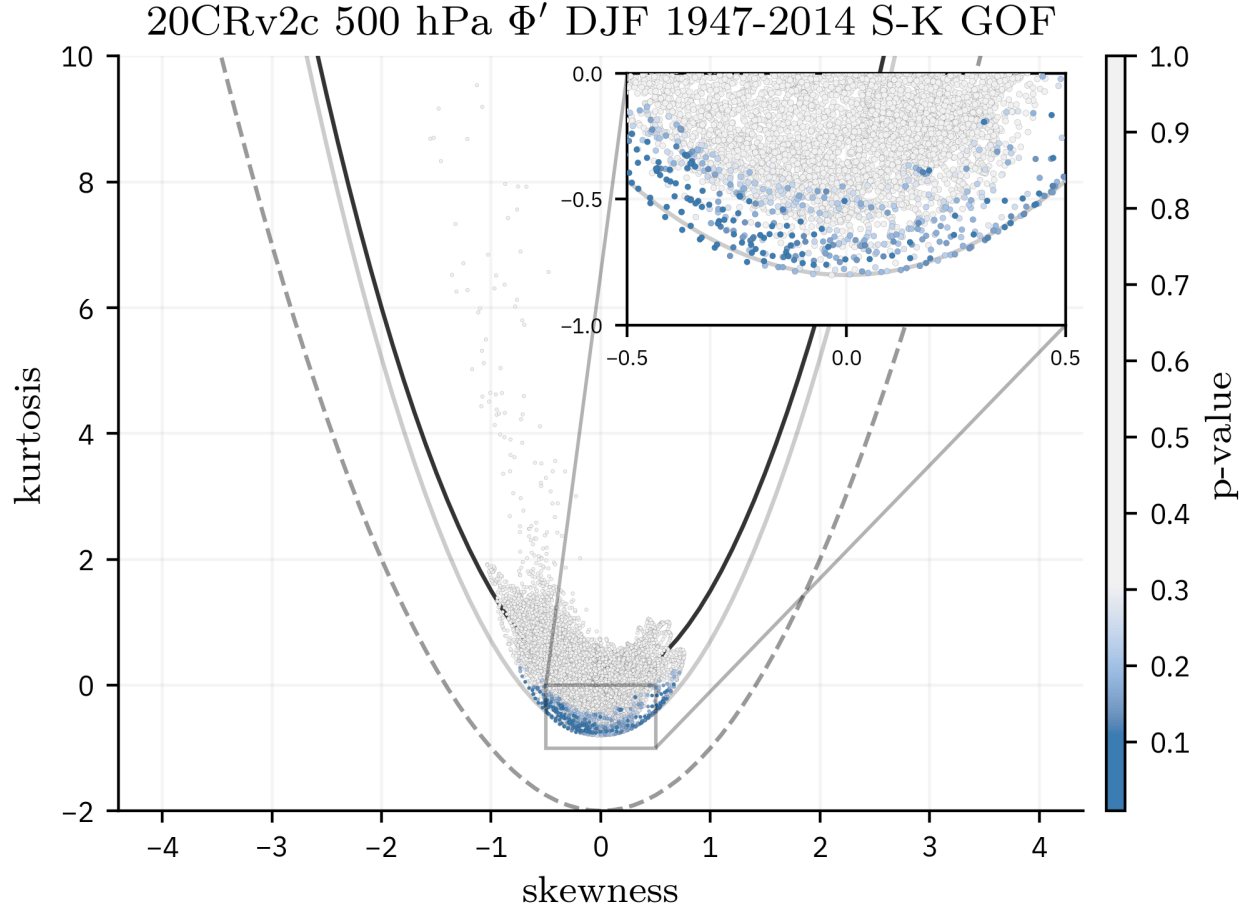


Figure 7.12: Skewness-kurtosis scatter plot showing the goodness-of-fit (GOF) p-value relative to the sample skewness and kurtosis computed for each time series of 500 hPa 20CRv2c geopotential height anomalies, 1947-2014. The solid black parabola indicates the skewness-kurtosis relationship $kurt \geq \left(\frac{3}{2}\right) skew^2$, where the light gray solid parabola is $kurt \geq \left(\frac{3}{2}\right) skew^2 - r$ where r is empirically fit to the data. The dotted gray line represents the $kurt = skew^2 - 2$ constraint.

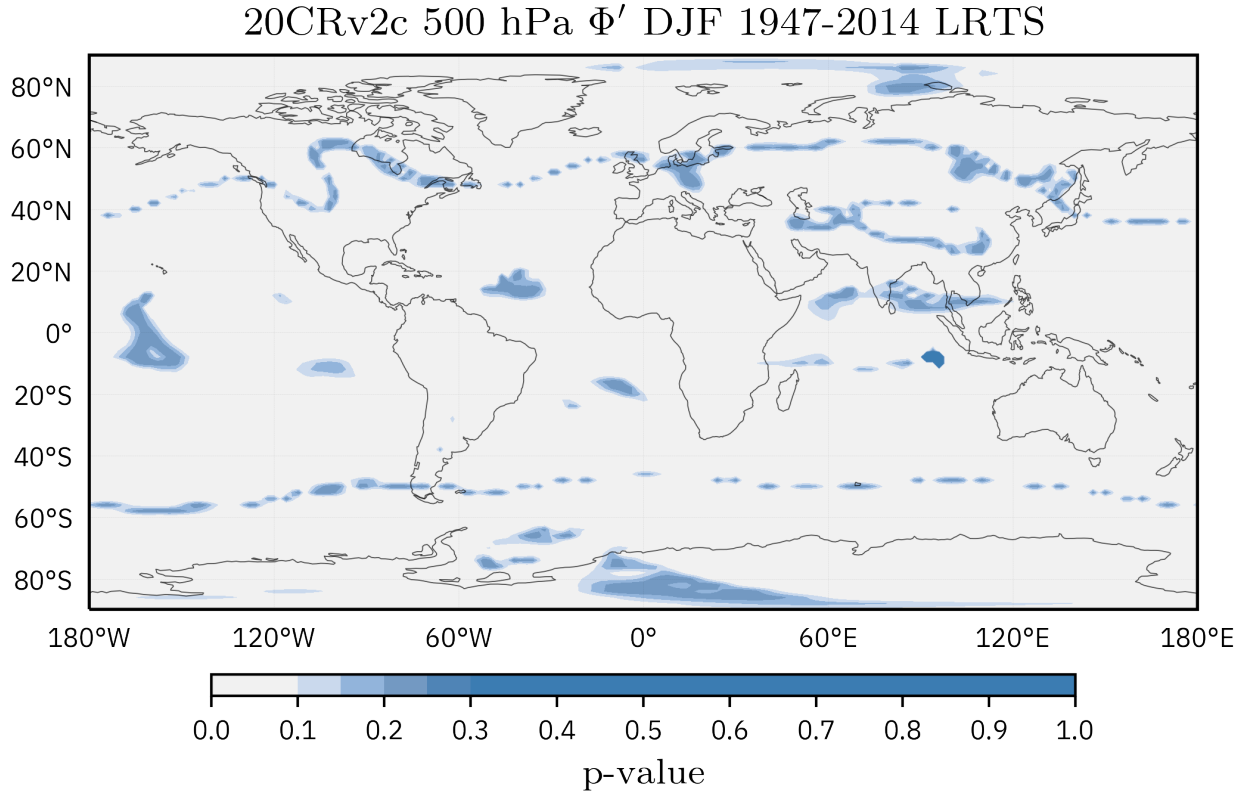


Figure 7.13: p-values of the likelihood ratio test statistic (LRTS) comparison between the estimated SGS and standard normal distributions for all gridded time series of standardized DJF 500 hPa 20CRv2c geopotential height anomalies, 1947-2014. Gray points correspond to grid points where the SGS distribution is statistically preferred to the standard normal distribution. Points gridded blue, especially those with values much greater than 0.1, correspond to a statistically insignificant preference for the SGS fit or a preference for the standard normal distribution.

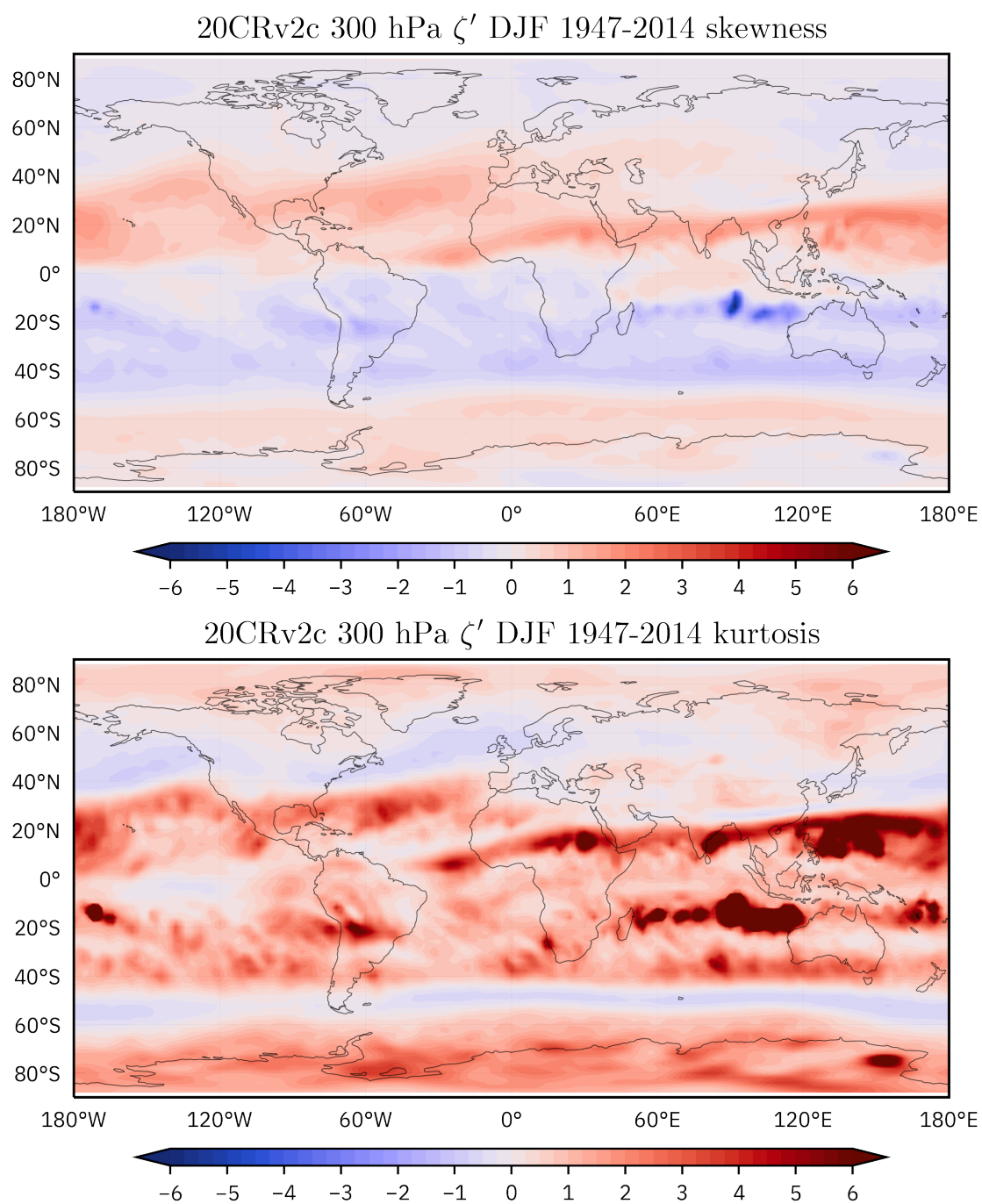


Figure 7.14: Maps of skewness and kurtosis of standardized DJF 300 hPa relative vorticity anomalies from the 20CRv2c dataset, 1947-2014.

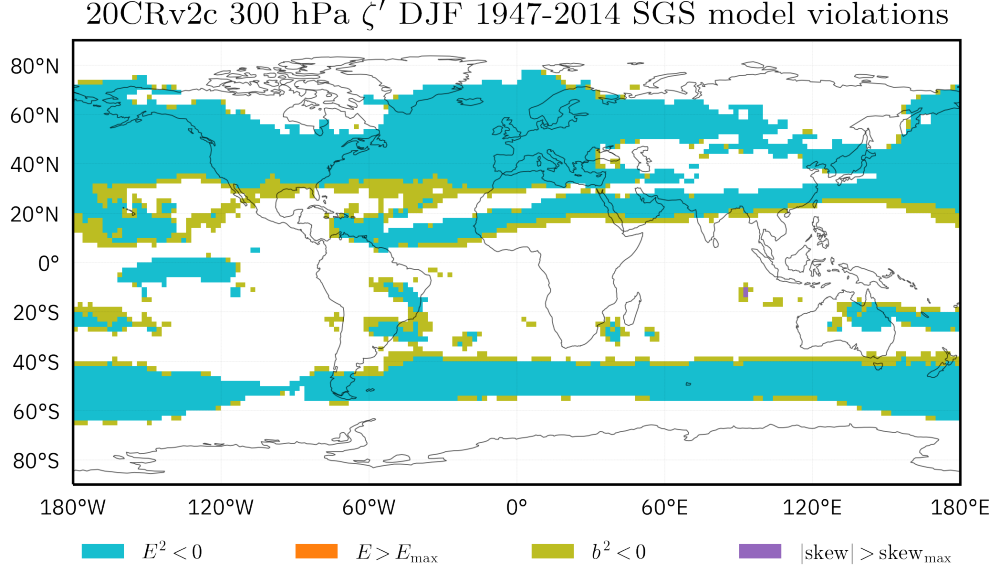


Figure 7.15: Map of the SGS model violations encountered when fitting an SGS distribution to time series of standardized DJF 300 hPa 20CRv2c relative vorticity anomalies, 1947-2014. Violations $E^2 < 0$ (cyan) and $b^2 < 0$ (olive) are most prominent.

decorrelation times are small ($\approx < 1$ week). As such, goodness-of-fit p-values of the relative vorticity anomalies do not require scrutiny as in the case of the air temperature and geopotential height anomalies. The p-value results from the goodness-of-fit tests are shown in Figure 7.17.

As the kurtosis of the 300 hPa relative vorticity anomalies (shown in Figure 7.14) is generally positive except in the regions of the mean storm track, Figure 7.17 indicates that the estimated SGS distribution is plausible over much of the globe. The major areas of implausibility include the mean location of the jet stream, as these areas are associated with regions of negative kurtosis. The goodness-of-fit test for the standard normal distribution shows similar patterns of implausibility, but performs best in near-Gaussian regions such as the band of near-zero skewness and kurtosis at 60°S latitude.

Figure 7.18 investigates further the relationship of skewness and kurtosis with the p-value. Moments calculated from time series of 300 hPa relative vorticity anomalies are not tightly clustered as the air temperature or geopotential height anomalies. There are also several points of implausible SGS distributions dispersed throughout the center of the parabola.

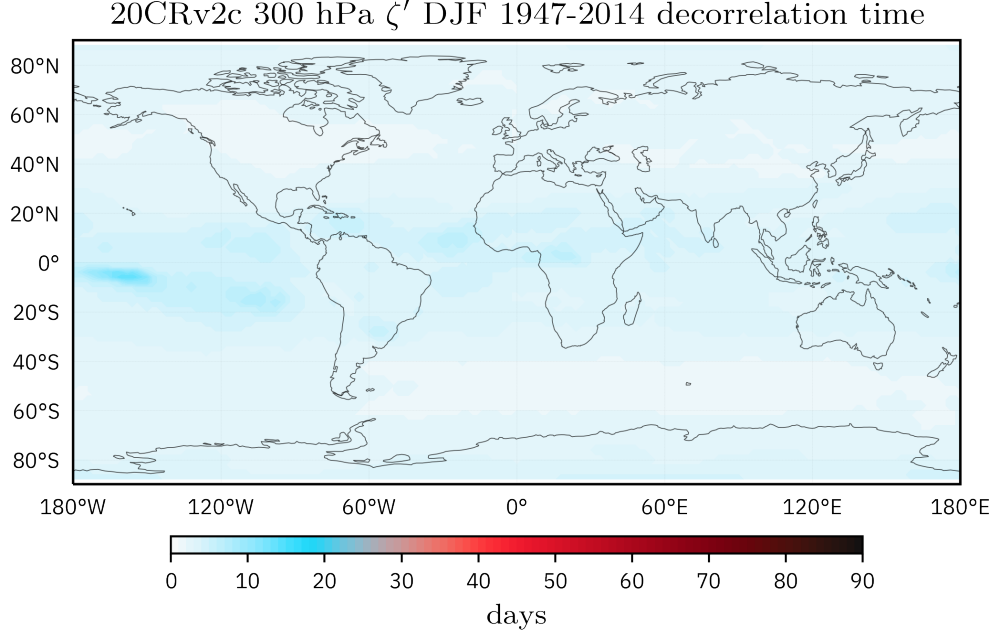


Figure 7.16: Decorrelation time scale in days computed using the $1/e$ autocorrelation threshold for the standardized DJF 300 hPa 20CRv2c relative vorticity anomalies, 1947-2014.

p-values obtained from the likelihood ratio test of the 950 hPa relative vorticity anomalies are given in Figure 7.19. Here, the estimated SGS distribution outperforms the fit of the standard normal distribution in all locations save those of near-zero skewness and kurtosis given in Figure 7.14. These near-Gaussian locations are the same mean storm tracks identified previously.

7.4 950 hPa zonal wind

Next, the 1947-2014 DJF global 950 hPa zonal wind anomalies are analyzed. After the SGS distribution is estimated at each grid point, the SGS model violations are tallied and given in Figure 7.21. Figure 7.21, like the previous variables, suggests that $E^2 < 0$ occurs in regions of negative kurtosis (see Figure 7.20). However, Figure 7.21 provides a somewhat clearer picture with respect to the $b^2 < 0$ violation, in that it appears to be primarily found in regions of positive kurtosis and a large magnitude of skewness. In other words, E is positive but not positive enough to satisfy the ϵ -constraint of Equation (4.10), which if satisfied ensures that E and g are positive.

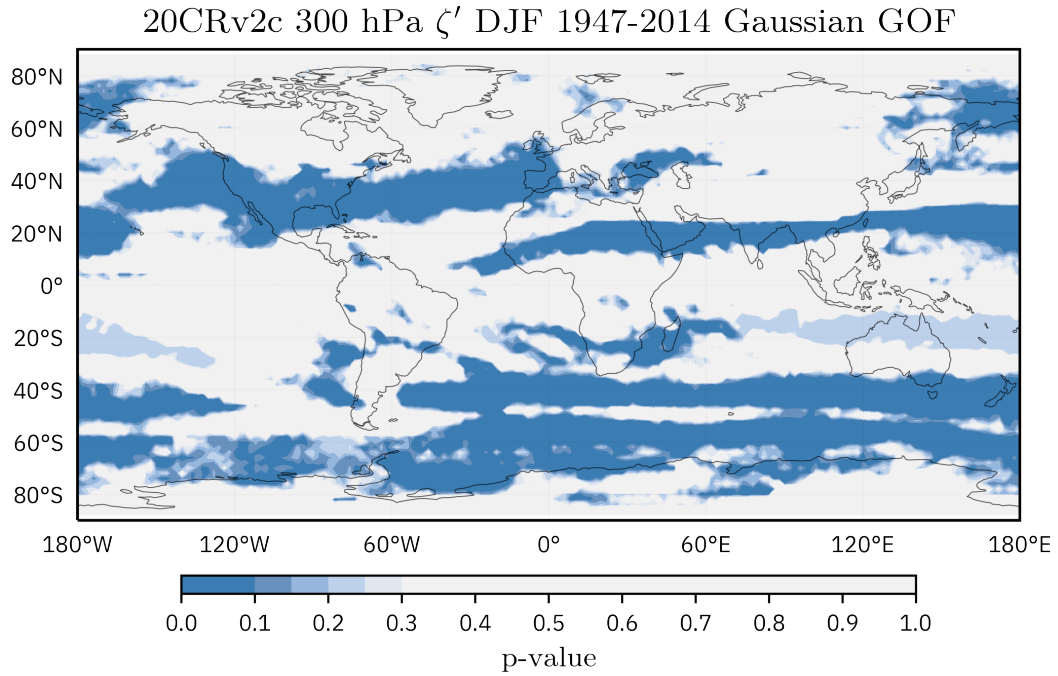
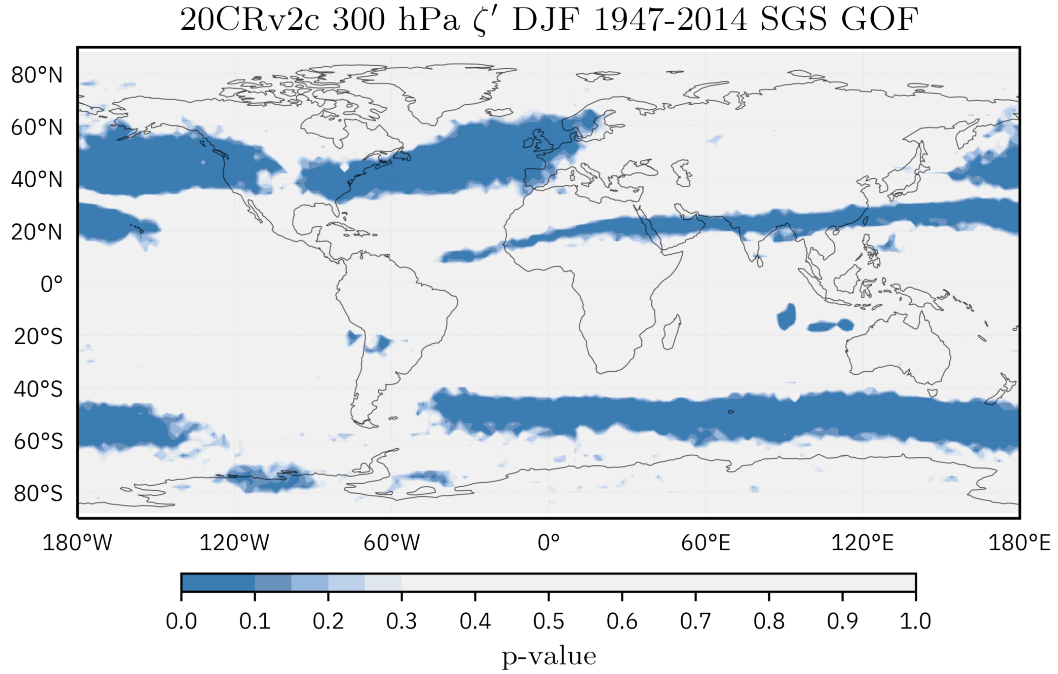


Figure 7.17: p-values for the estimated SGS (top) and standard normal (bottom) goodness-of-fit (GOF) test computed using 100 bootstrap iterations for all gridded time series of standardized DJF 300 hPa 20CRv2c relative vorticity anomalies, 1947-2014. Gray points correspond to a plausible SGS fit, while points gridded blue, especially those with values less than 0.1, correspond to an implausible SGS fit.

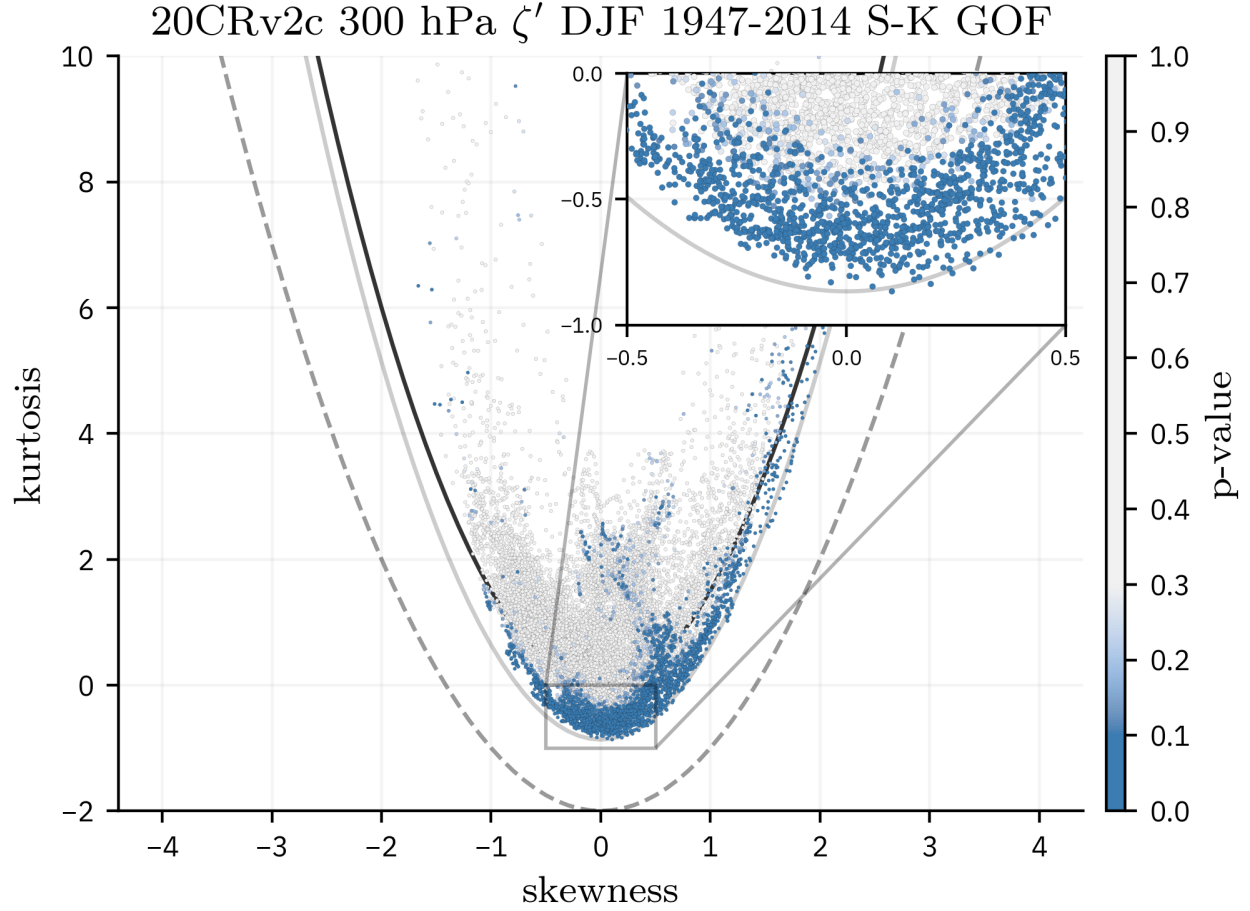


Figure 7.18: Skewness-kurtosis scatter plot showing the goodness-of-fit (GOF) p-value relative to the sample skewness and kurtosis computed for each time series of 300 hPa 20CRv2c relative vorticity anomalies, 1947-2014. The solid black parabola indicates the skewness-kurtosis relationship $kurt \geq \left(\frac{3}{2}\right) skew^2$, where the light gray solid parabola is $kurt \geq \left(\frac{3}{2}\right) skew^2 - r$ where r is empirically fit to the data. The dotted gray line represents the $kurt = skew^2 - 2$ constraint.

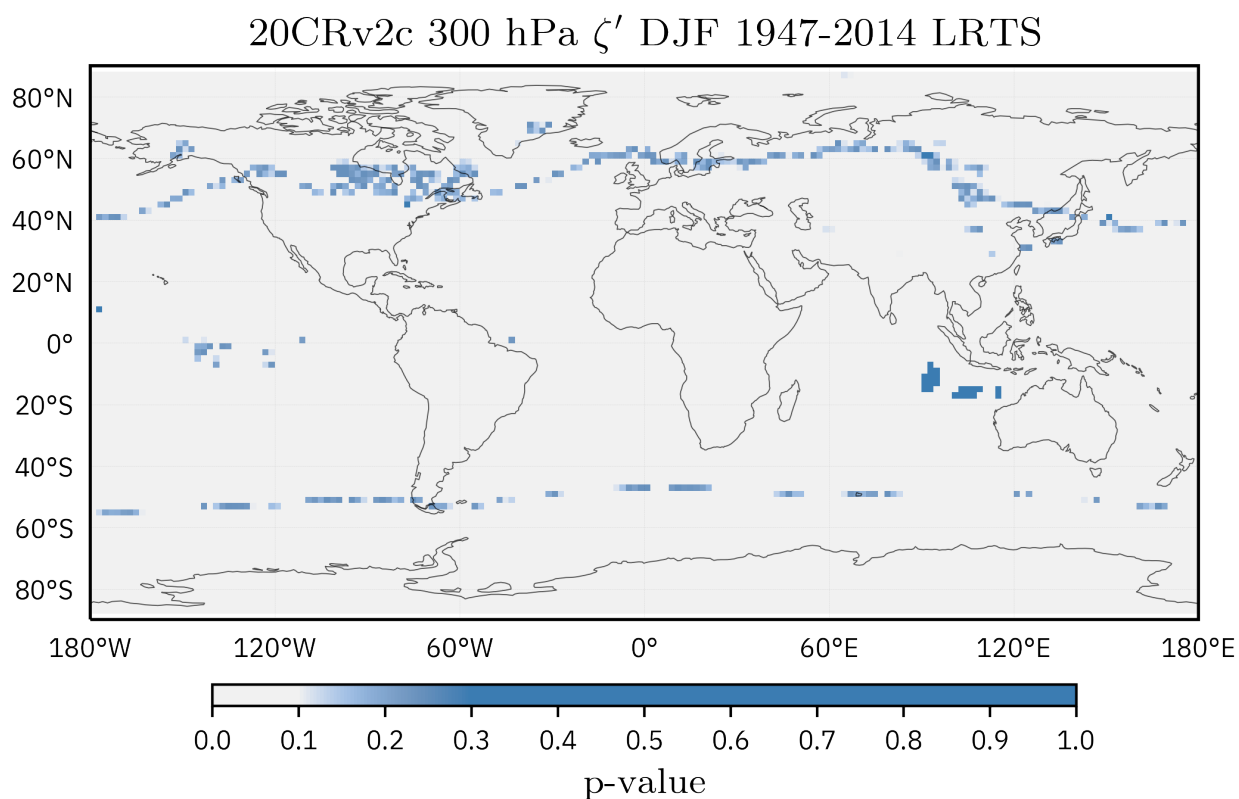


Figure 7.19: p-values of the likelihood ratio test statistic (LRTS) comparison between the estimated SGS and standard normal distributions for all gridded time series of standardized DJF 300 hPa 20CRv2c relative vorticity anomalies, 1947-2014. Gray points correspond to grid points where the SGS distribution is statistically preferred to the standard normal distribution. Points gridded blue, especially those with values much greater than 0.1, correspond to a statistically insignificant preference for the SGS fit or a preference for the standard normal distribution.

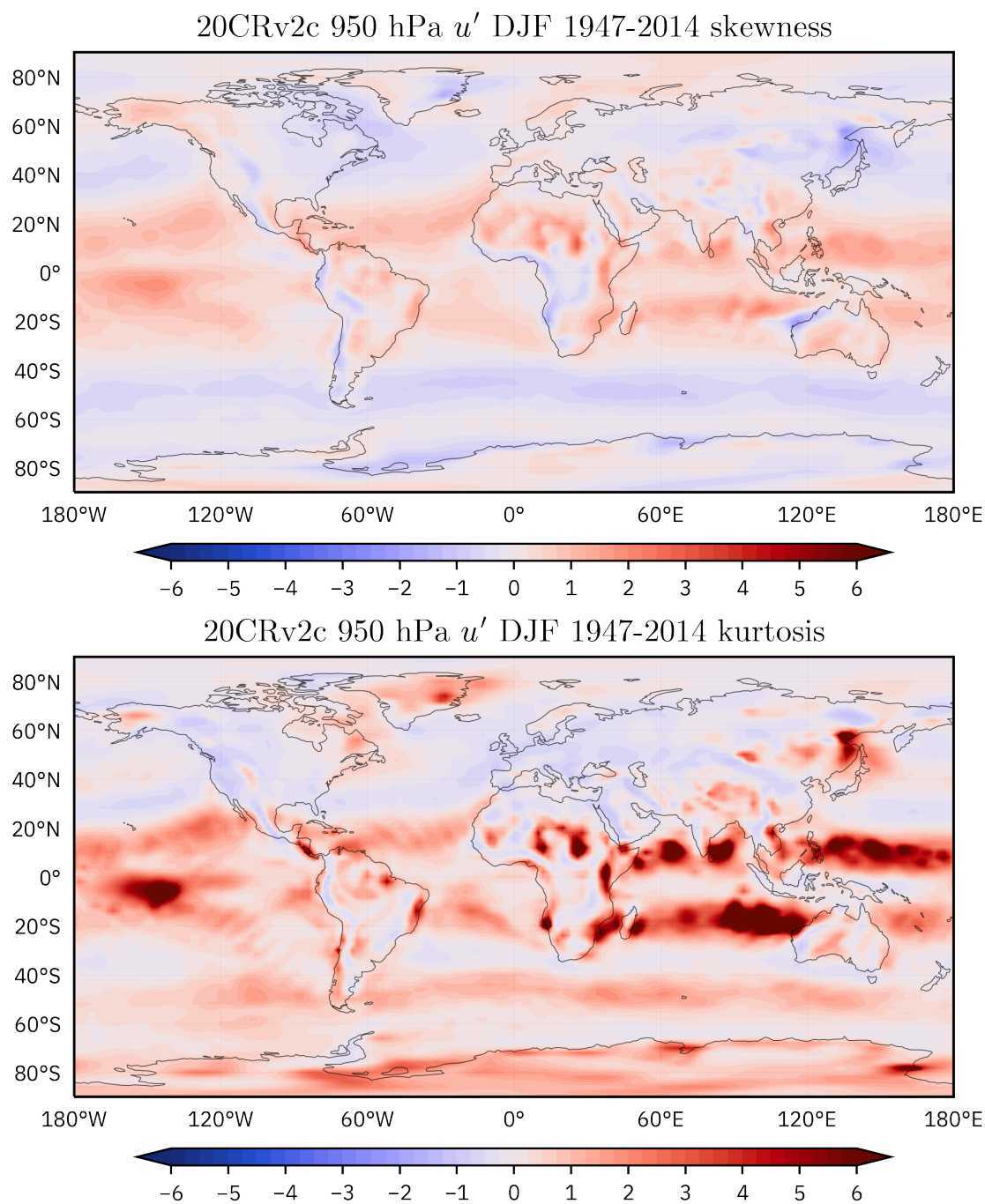


Figure 7.20: Maps of skewness and kurtosis of standardized DJF 950 hPa zonal wind anomalies from the 20CRv2c dataset, 1947-2014.

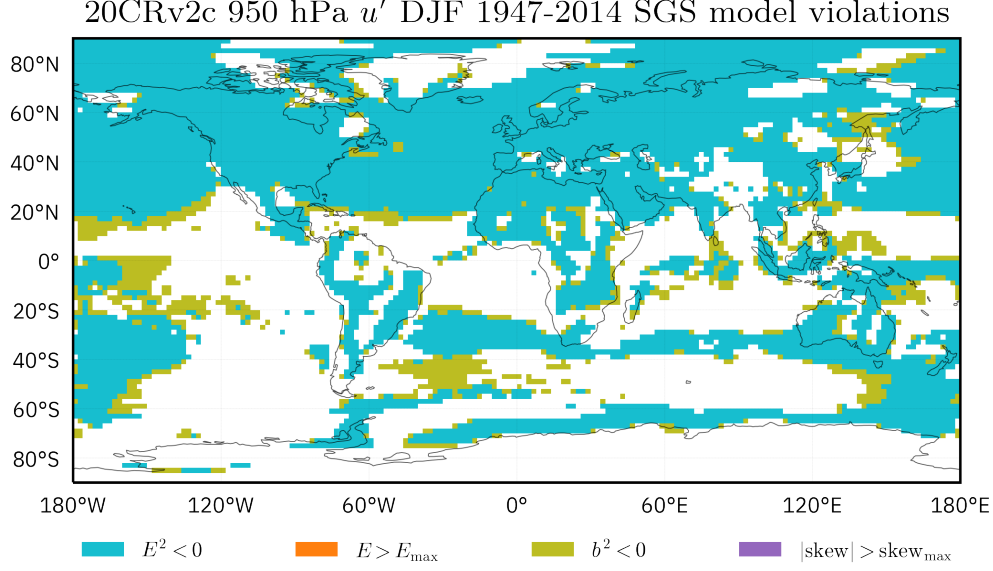


Figure 7.21: Map of the SGS model violations encountered when fitting an SGS distribution to time series of standardized DJF 950 hPa 20CRv2c zonal wind anomalies, 1947-2014. Violations $E^2 < 0$ (cyan) and $b^2 < 0$ (olive) are most prominent.

The decorrelation time scales are again computed from the time series of anomalies at each grid point. Figure 7.22 shows the decorrelation time scales for all 950 hPa zonal wind anomalies.

In contrast to the decorrelation time scales of both the air temperature and geopotential height anomalies analyzed previously, there is no exceptionally long decorrelation times present for the 950 hPa zonal wind, suggesting that wind anomalies are not as persistent and typically rapidly decorrelate. The most significant region of large decorrelation times is the western equatorial Pacific, where lags of three weeks are found. p-value results from the goodness-of-fit tests are shown in Figure 7.23.

With smaller decorrelation times in the tropics, there is a clearer picture of the significance of the goodness-of-fit for the estimated SGS distributions there. Overall, however, a similar narrative emerges, as regions of near-Gaussianity tend to produce plausible fits to the standard normal distribution while regions that have larger magnitudes of skewness and positive kurtosis produce more plausible SGS fits (see Figure 7.20 for maps of skewness and kurtosis). Figure 7.24 provides a more nuanced look at the values of skewness and kurtosis that make up the set of plausible SGS distributions. While many implausible SGS distributions are fit to time series that violate the

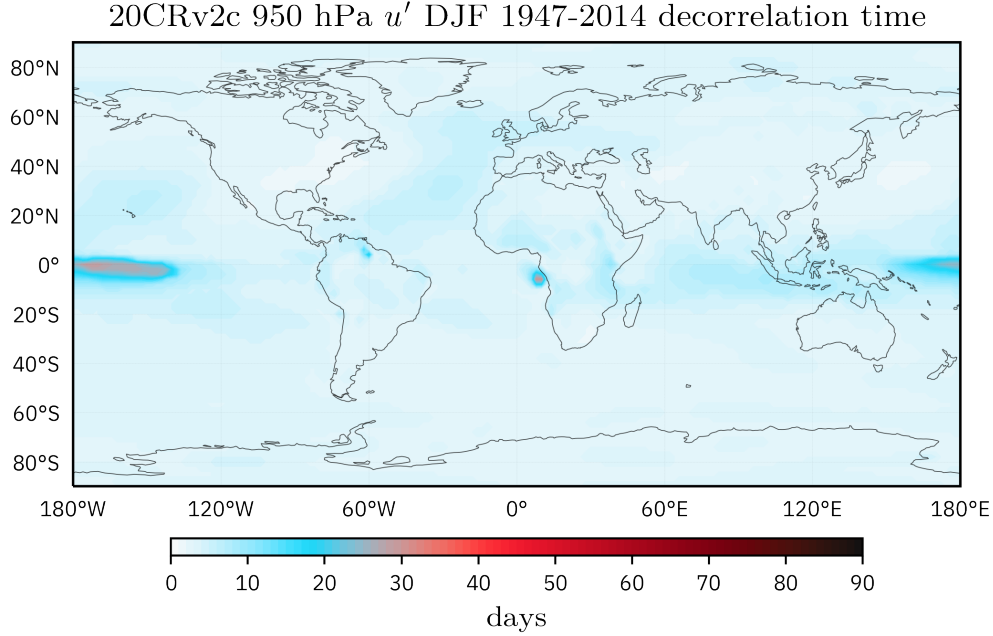


Figure 7.22: Decorrelation time scale in days computed using the $1/e$ autocorrelation threshold for the standardized DJF 950 hPa 20CRv2c zonal wind anomalies, 1947-2014.

skewness-kurtosis constraint of Equation (4.9), there are several time series which have very large values of skewness and/or kurtosis that do not result in a plausible SGS distribution. The spread of the skewness-kurtosis scatter points in Figure 7.24 is also substantially larger than in the 950 hPa air temperature and 500 hPa geopotential height anomalies.

The implausibility of the SGS distribution over many regions of the northern hemisphere, most especially over the continents, is related to the negative kurtosis endemic there. However, the implausible standard normal distributions are typically located in regions of large skewness, large kurtosis or both.

The p-values of the likelihood ratio test comparing the estimated SGS distribution to the standard normal distribution are shown in Figure 7.25. The grid points where the estimated SGS distribution compares unfavorably to the standard normal distribution are given in blue, which upon comparison to the skewness and kurtosis maps of Figure 7.20 are all located in regions of near-zero skewness and zero to negative kurtosis.

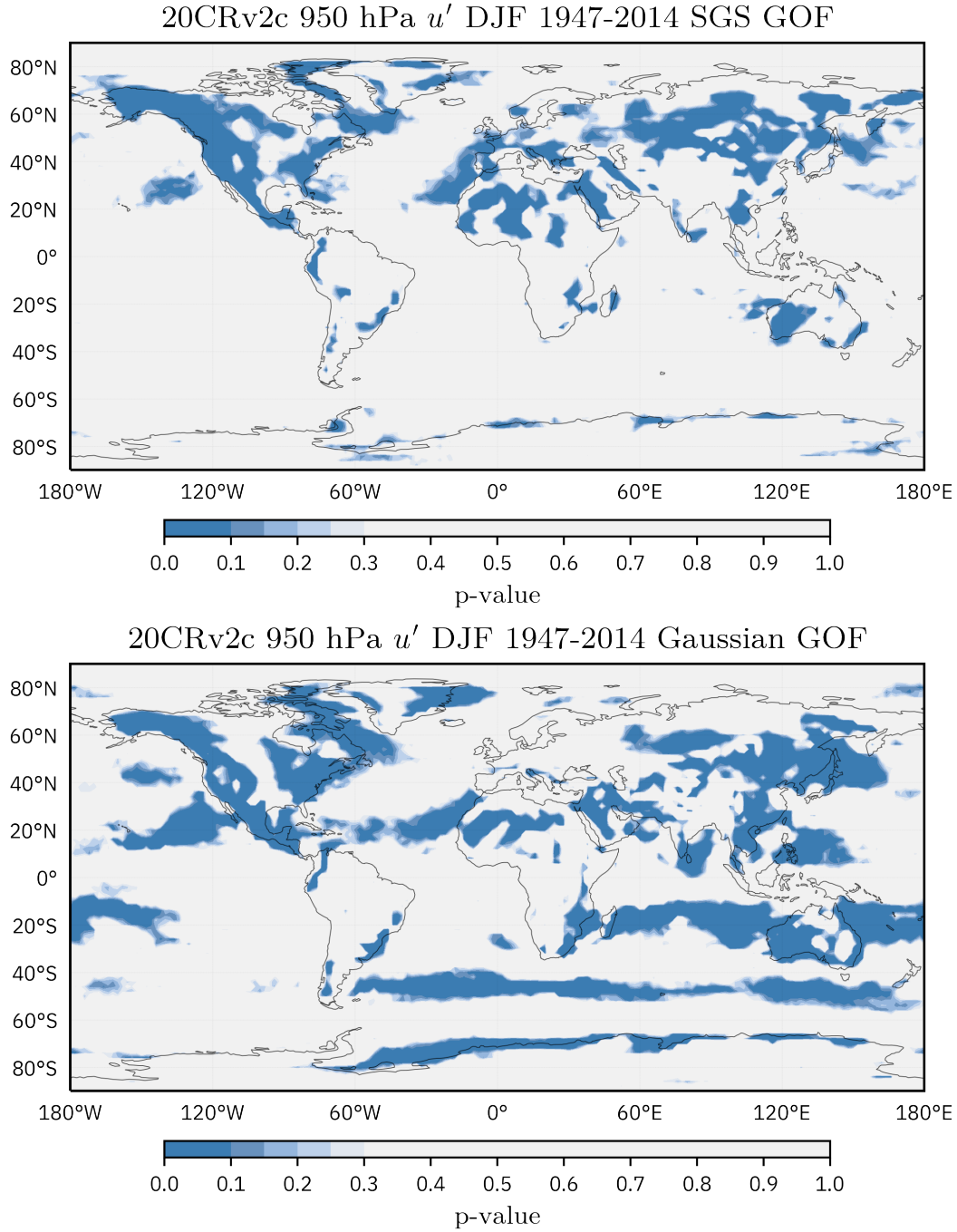


Figure 7.23: p-values for the estimated SGS (top) and standard normal (bottom) goodness-of-fit (GOF) test computed using 100 bootstrap iterations for all gridded time series of standardized DJF 950 hPa 20CRv2c zonal wind anomalies, 1947-2014. Gray points correspond to a plausible SGS fit, while points gridded blue, especially those with values less than 0.1, correspond to an implausible SGS fit.

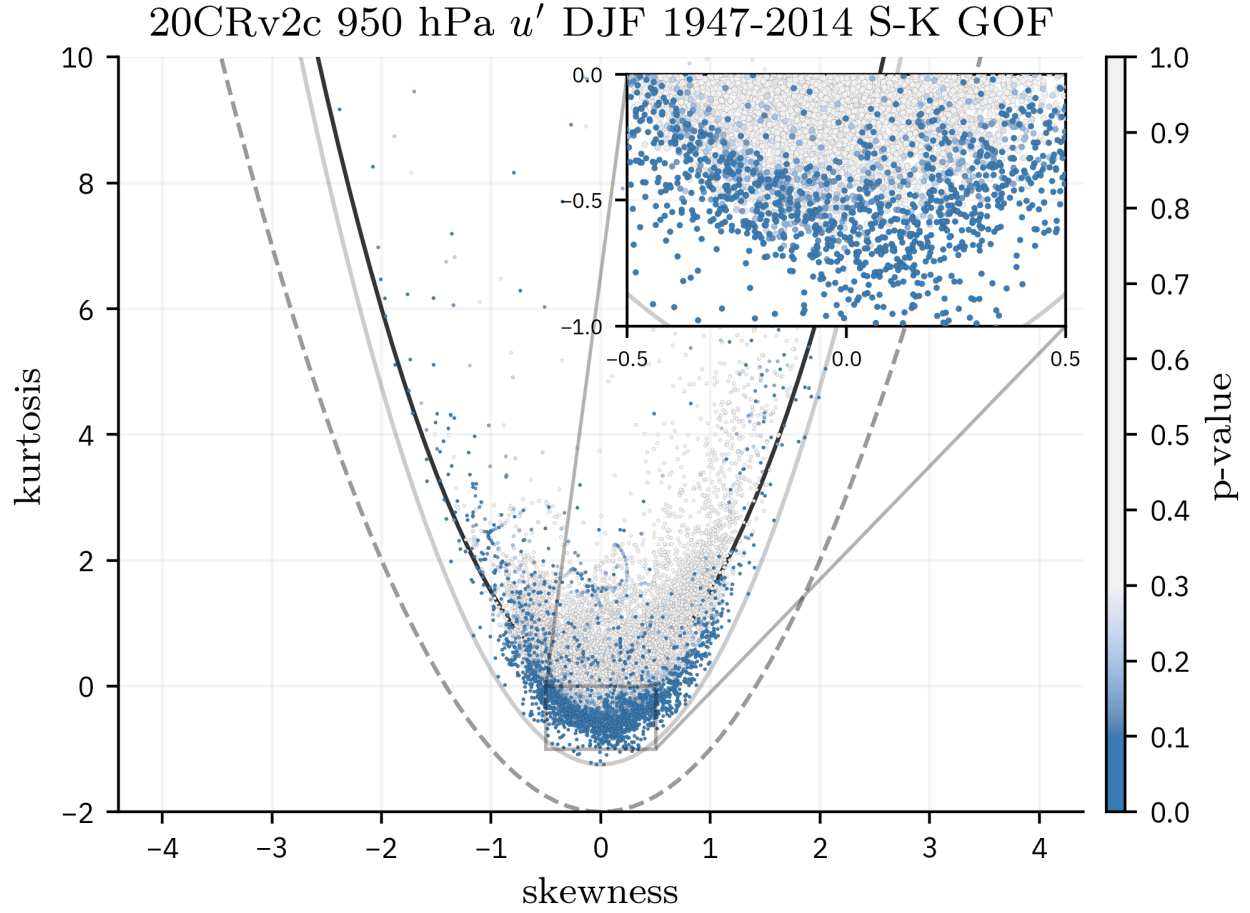


Figure 7.24: Skewness-kurtosis scatter plot showing the goodness-of-fit (GOF) p-value relative to the sample skewness and kurtosis computed for each time series of 950 hPa 20CRv2c zonal wind anomalies, 1947-2014. The solid black parabola indicates the skewness-kurtosis relationship $\text{kurt} \geq \left(\frac{3}{2}\right) \text{skew}^2$, where the light gray solid parabola is $\text{kurt} \geq \left(\frac{3}{2}\right) \text{skew}^2 - r$ where r is empirically fit to the data. The dotted gray line represents the $\text{kurt} = \text{skew}^2 - 2$ constraint.

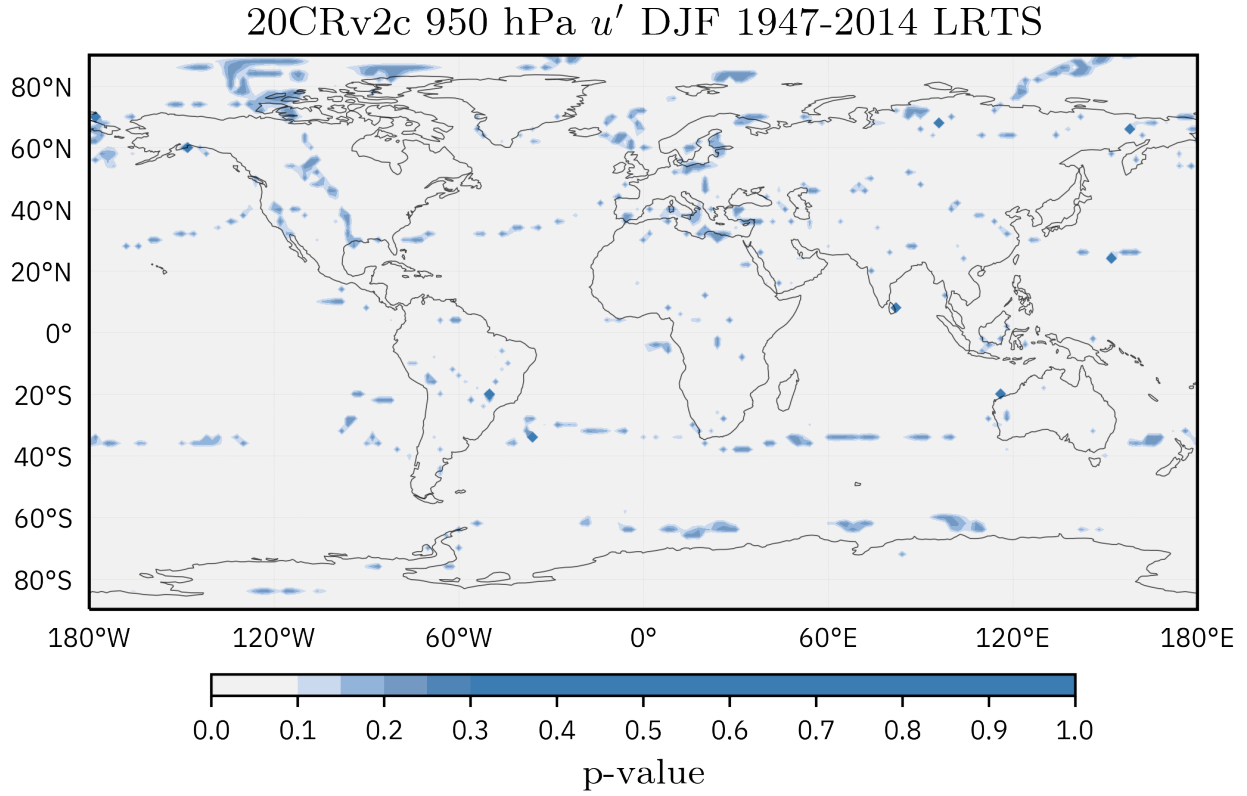


Figure 7.25: p-values of the likelihood ratio test statistic (LRTS) comparison between the estimated SGS and standard normal distributions for all gridded time series of standardized DJF 950 hPa 20CRv2c zonal wind anomalies, 1947-2014. Gray points correspond to grid points where the SGS distribution is statistically preferred to the standard normal distribution. Points gridded blue, especially those with values much greater than 0.1, correspond to a statistically insignificant preference for the SGS fit or a preference for the standard normal distribution.

7.5 950 hPa meridional wind

After the SGS distribution is fit to each gridded time series of 950 hPa meridional wind anomalies, the SGS model violations are compiled and compared with maps of skewness and kurtosis, given in Figure 7.26. The SGS violations given in Figure 7.27 are similar to those of the 950 hPa zonal wind anomalies, where violations of $E^2 < 0$ are associated with regions of negative kurtosis and violations of $b^2 < 0$ are found in regions of positive kurtosis and larger magnitude of skewness.

Figure 7.28 shows the decorrelation time scales for all time series of 950 hPa meridional wind anomalies. Like those of the 950 hPa zonal wind anomalies, the decorrelation times are relatively small (≈ 1 week) except for a few isolated points in tropical Africa and a strip of the equatorial Pacific Ocean. Again, while points of large decorrelation time scales should evoke a sense of caution in interpreting large goodness-of-fit p-values as overwhelmingly plausible, the meridional wind anomalies here do not suffer the large latitudinal regions of decorrelation times as the air temperature and geopotential anomalies did.

After the decorrelation times of Figure 7.28 are computed, p-value results from the goodness-of-fit tests are shown in Figure 7.29.

The goodness-of-fit values for the SGS distribution and the standard normal distribution tabulated in Figure 7.29 are best compared with the skewness and kurtosis maps of the meridional wind anomalies, given in Figure 7.26. Even as the sign of the 950 hPa meridional wind skewness has a hemispherical dependency reflecting the general circulation of the atmosphere, it is the magnitude and sign of the kurtosis that determines the goodness-of-fit for both the estimated SGS and the standard normal cases. As expected, the standard normal fit is poor in regions of large skewness and kurtosis, while the SGS fit struggles again in regions of negative skewness. Figure 7.30 looks at the skewness-kurtosis relationship in more detail. Like that of the 950 hPa zonal wind anomalies, the primary region of implausibility is located below the $K = 3/2S^2$ curve, shown in black in Figure 7.30. However, the moments calculated from the time series of anomalies are not as tightly clustered as the air temperature or geopotential height anomalies and there are several implausible SGS distributions dispersed throughout the parabola of points.

p-values computed from the likelihood ratio test for the 950 hPa meridional wind anomalies are given in Figure 7.31. Again, the estimated SGS distribution outperforms the fit of the standard normal distribution in all but the most Gaussian of points. An exception to this is found in the

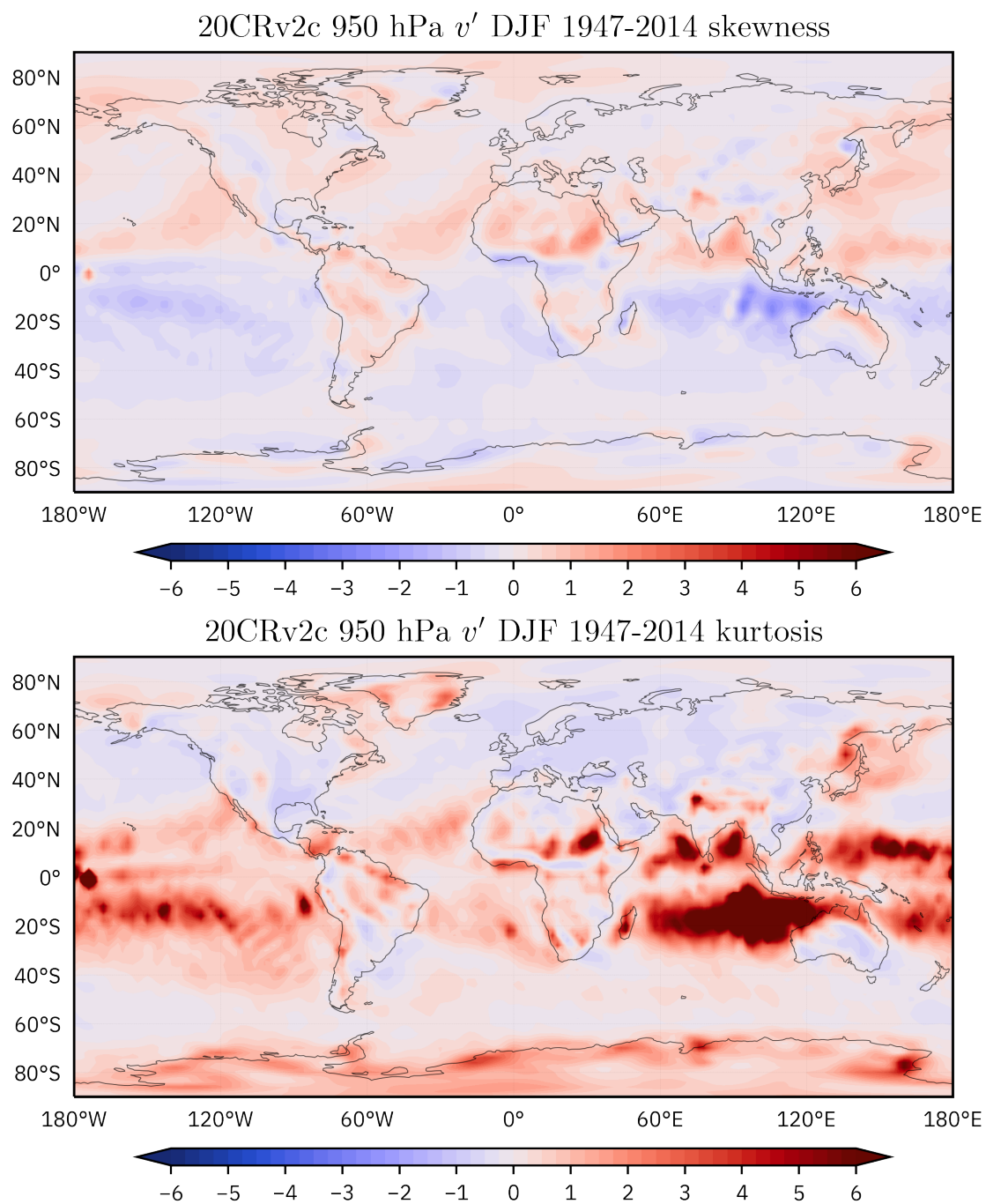


Figure 7.26: Maps of skewness and kurtosis of standardized DJF 950 hPa meridional wind anomalies from the 20CRv2c dataset, 1947-2014.

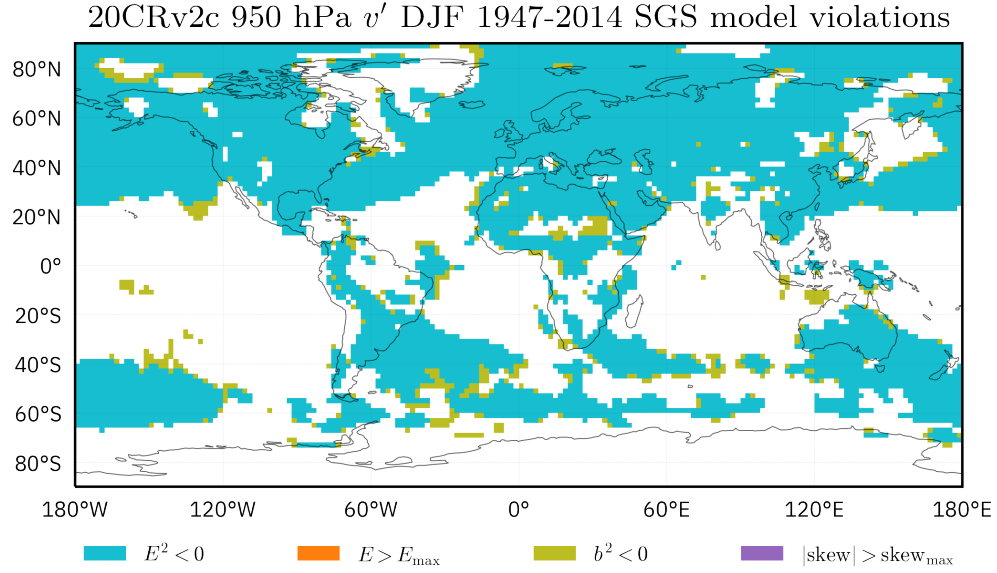


Figure 7.27: Map of the SGS model violations encountered when fitting an SGS distribution to time series of standardized DJF 950 hPa 20CRv2c meridional wind anomalies, 1947-2014. Violations $E^2 < 0$ (cyan) and $b^2 < 0$ (olive) are most prominent.

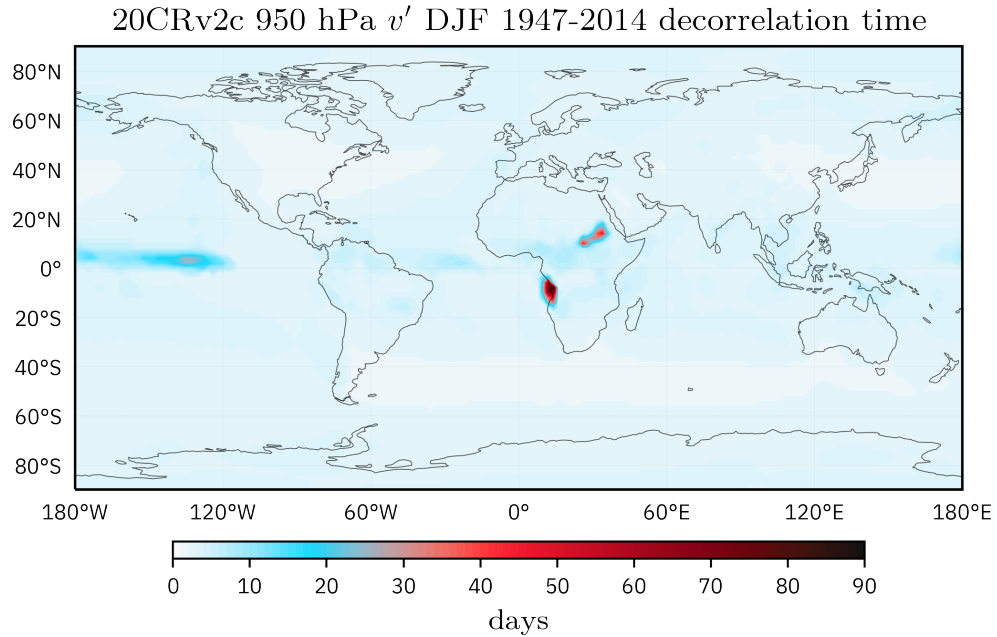


Figure 7.28: Decorrelation time scale in days computed using the $1/e$ autocorrelation threshold for the standardized DJF 950 hPa 20CRv2c meridional wind anomalies, 1947-2014.

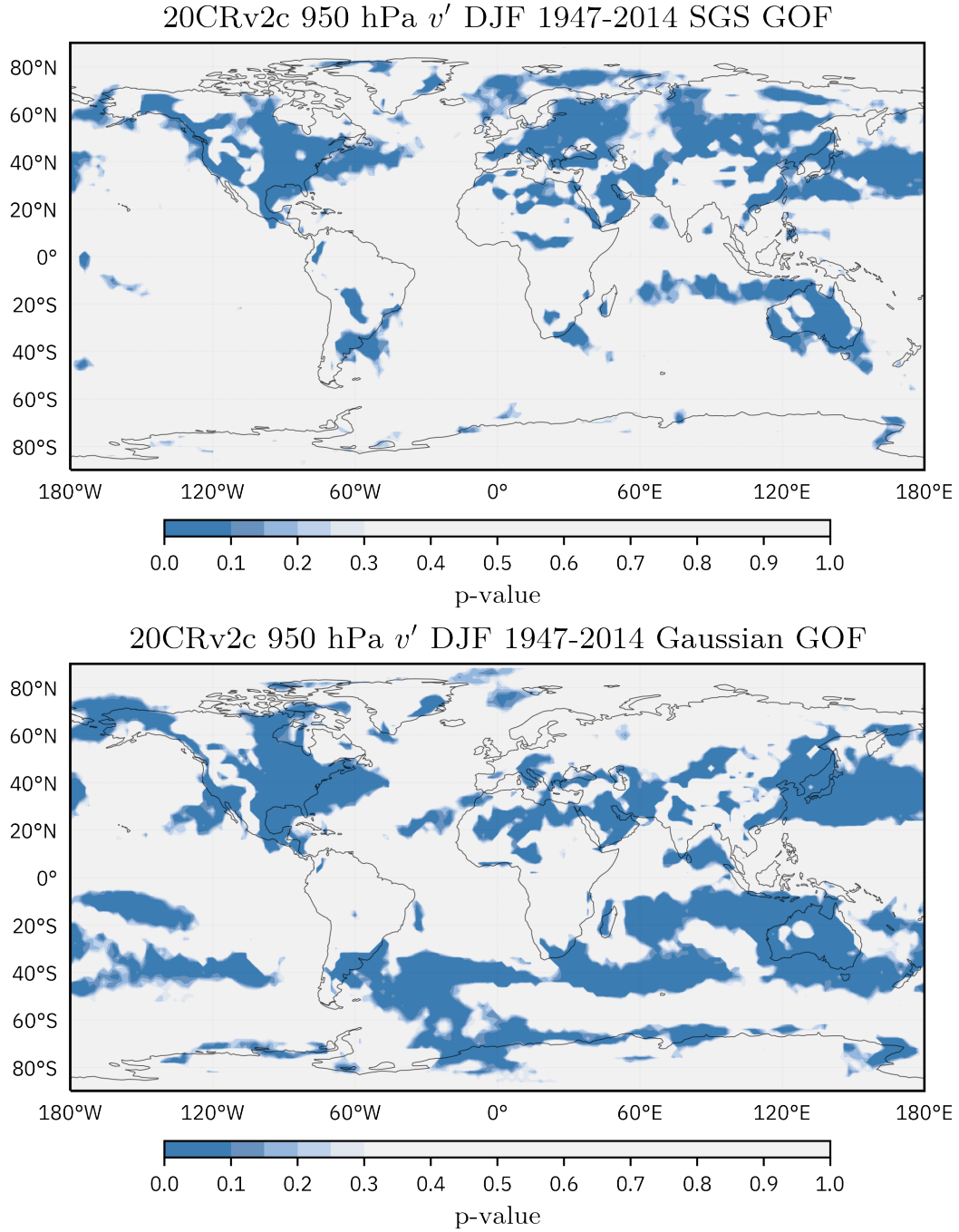


Figure 7.29: p-values for the estimated SGS (top) and standard normal (bottom) goodness-of-fit (GOF) test computed using 100 bootstrap iterations for all gridded time series of standardized DJF 950 hPa 20CRv2c meridional wind anomalies, 1947-2014. Gray points correspond to a plausible SGS fit, while points gridded blue, especially those with values less than 0.1, correspond to an implausible SGS fit.

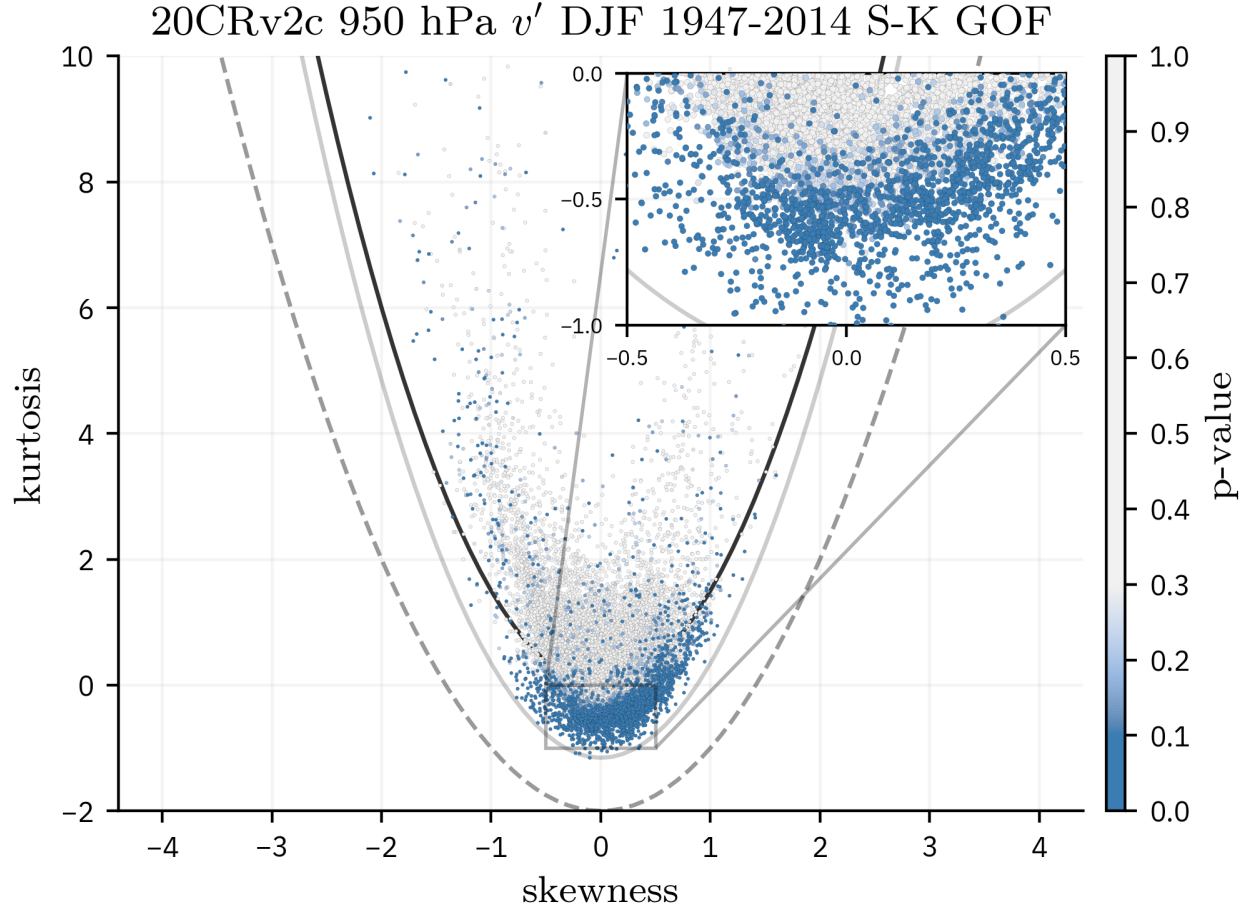


Figure 7.30: Skewness-kurtosis scatter plot showing the goodness-of-fit (GOF) p-value relative to the sample skewness and kurtosis computed for each time series of 950 hPa 20CRv2c meridional wind anomalies, 1947-2014. The solid black parabola indicates the skewness-kurtosis relationship $kurt \geq \left(\frac{3}{2}\right) skew^2$, where the light gray solid parabola is $kurt \geq \left(\frac{3}{2}\right) skew^2 - r$ where r is empirically fit to the data. The dotted gray line represents the $kurt = skew^2 - 2$ constraint.

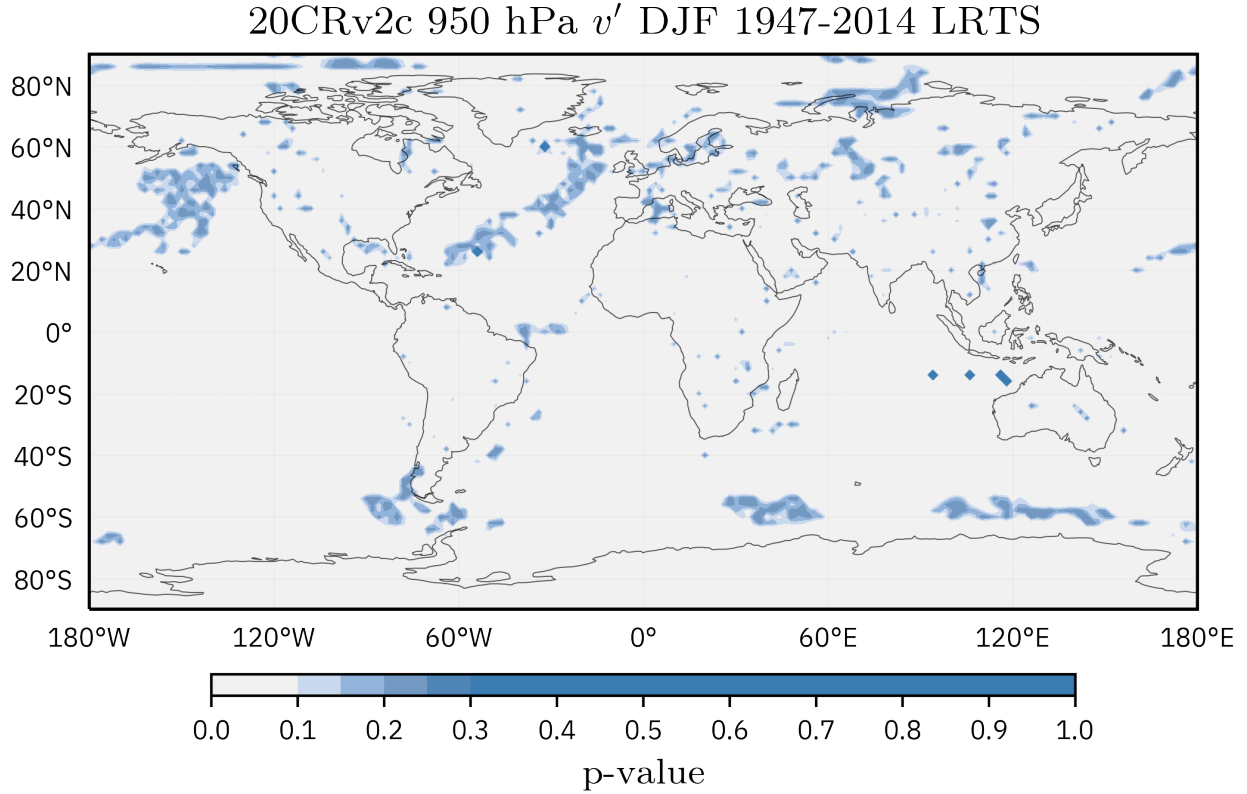


Figure 7.31: p-values of the likelihood ratio test statistic (LRTS) comparison between the estimated SGS and standard normal distributions for all gridded time series of standardized DJF 950 hPa 20CRv2c meridional wind anomalies, 1947-2014. Gray points correspond to grid points where the SGS distribution is statistically preferred to the standard normal distribution. Points gridded blue, especially those with values much greater than 0.1, correspond to a statistically insignificant preference for the SGS fit or a preference for the standard normal distribution.

extreme southeast Indian Ocean, where four points located in a region of very large positive kurtosis do not significantly favor the SGS distribution estimated there. More analysis is needed to isolate the cause of these few highly non-Gaussian points that do not attain a significant level of plausibility.

7.6 500 hPa vertical velocity

Finally, the SGS distribution is fit to time series of 500 hPa vertical velocity anomalies. The violations of the SGS model produced during the fit are given in Figure 7.33. As the kurtosis of the

500 hPa vertical velocity anomalies is overwhelmingly positive, the more prevalent $E^2 < 0$ violation is found sparingly. However, a large portion of the equator was subjected to the $b^2 < 0$ violation, and a region in the southeastern Indian Ocean contains skewness values that are larger than the $\sqrt{32}$ bound set while the kurtosis exists.

Figure 7.34 shows the decorrelation time scales for all 500 hPa vertical velocity anomalies. The 500 hPa vertical velocity anomalies are generally devoid of any extremely long decorrelation times, save a region in the equatorial Pacific that exhibits decorrelation times of approximately one month. Undertaking the goodness-of-fit test for the SGS and standard normal distributions produces the p-values given in Figure 7.35.

The maps of the goodness-of-fit p-values computed for the 500 hPa vertical velocity are significantly different than any of the prior variables analyzed thus far. This is a reflection of the skewness and kurtosis maps of the 500 hPa vertical velocity anomalies, given in Figure 7.32, which are highly asymmetric. Skewness values of the vertical velocity are overwhelmingly negative, especially at the equator and ITCZ, indicative of the vertical motion due to convection in those regions. Meanwhile, the kurtosis is positive nearly everywhere, though most strongly in ocean basins immediately surrounding the equator. p-values of the standard normal fit in Figure 7.32 are nearly entirely implausible, except for regions where the skewness is near-zero and the positive kurtosis is weakest. In contrast, the estimated SGS distributions produce plausible fits in regions of moderate positive kurtosis, but struggles in regions of severe positive kurtosis and near-zero kurtosis.

Figure 7.38 shows the skewness-kurtosis relationship alongside the p-value computed from the goodness-of-fit test. It is clear that the overwhelmingly negative skewness and positive kurtosis of many grid points contributes greatly to the implausibility of the SGS distribution there. However, it is curious that skewness and kurtosis values that meet the skewness-kurtosis constraint of Equation (4.9) produce such a poor fit. The pdf and ECDF of one of these points, located at 20°N, 150°W, is shown in Figures 7.36 and 7.37, respectively. Note the large peakedness and heavy, negative tail associated with the negative skewness and large positive kurtosis. Also, as the time series located at 20°N, 150°W only has a decorrelation time of three days, the KS statistics calculated from the bootstrapping goodness-of-fit test will be highly competitive with the KS statistic calculated from the observations. As such, it appears that the SGS fit methodology, which uses the method-of-

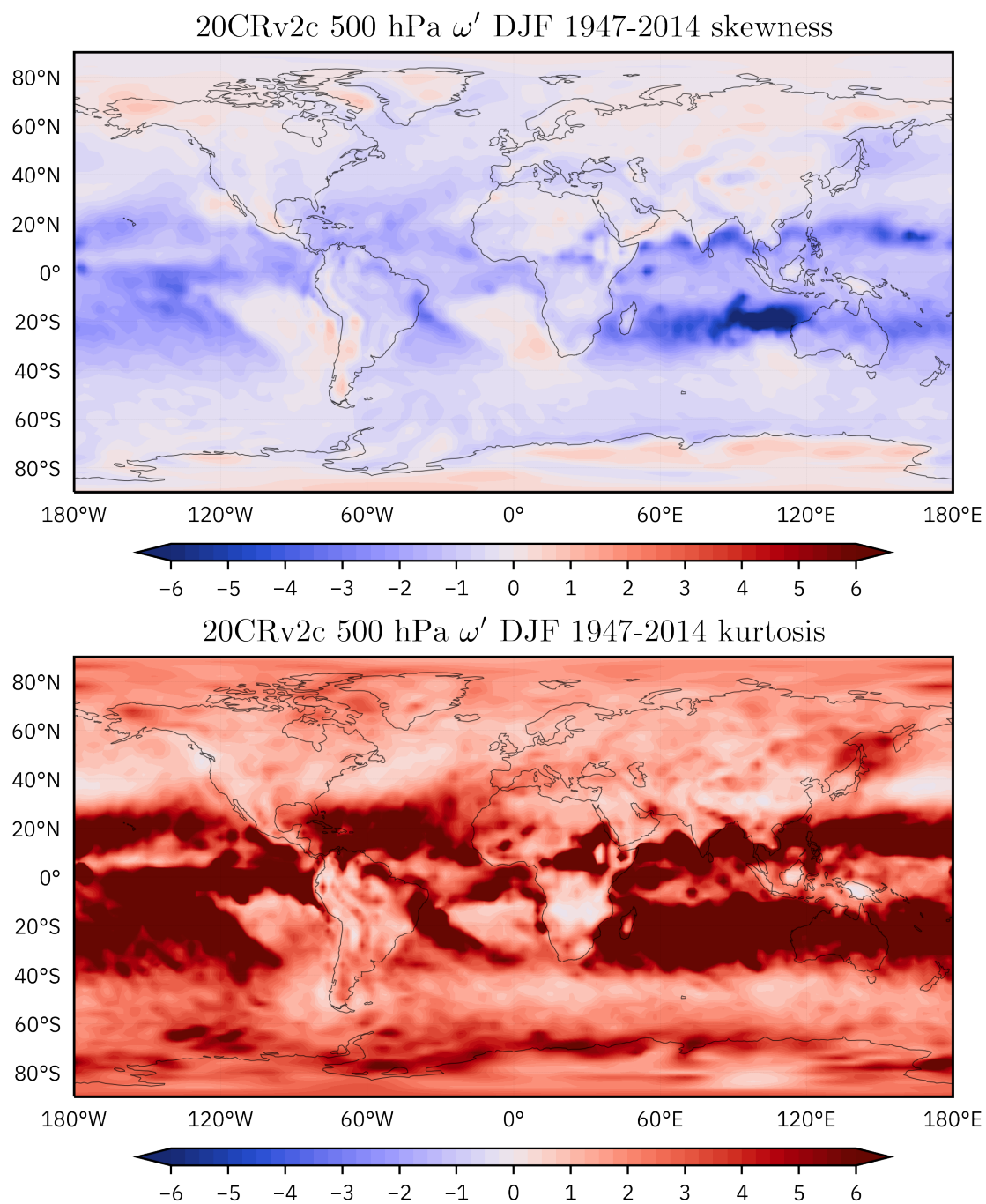


Figure 7.32: Maps of skewness and kurtosis of standardized DJF 500 hPa vertical velocity anomalies from the 20CRv2c dataset, 1947-2014.

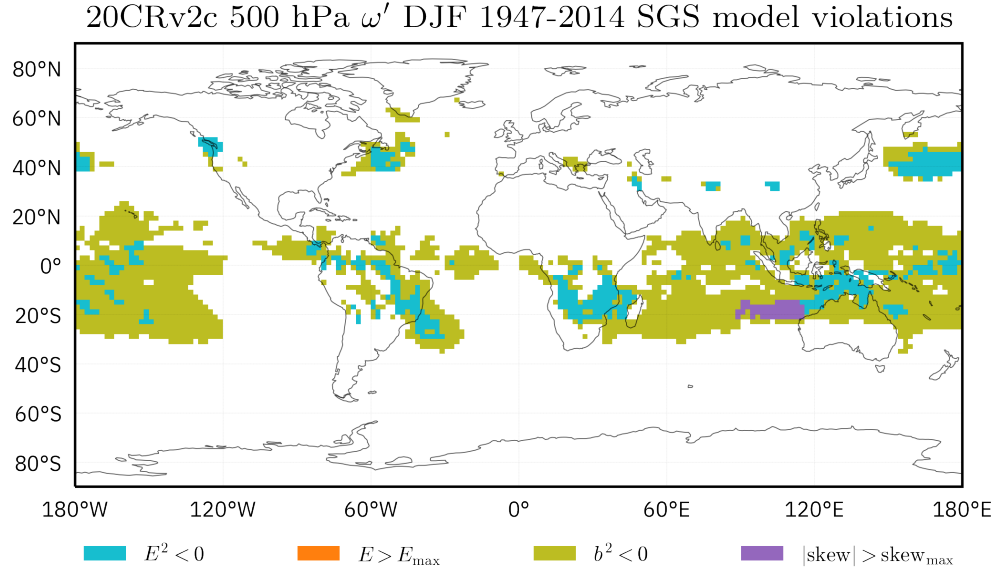


Figure 7.33: Map of the SGS model violations encountered when fitting an SGS distribution to time series of standardized DJF 500 hPa 20CRv2c vertical velocity anomalies, 1947-2014. Violations $E^2 < 0$ (cyan), $b^2 < 0$ (olive) and $|\text{skew}| > \text{skew}_{\max}$ (purple) are prominent.

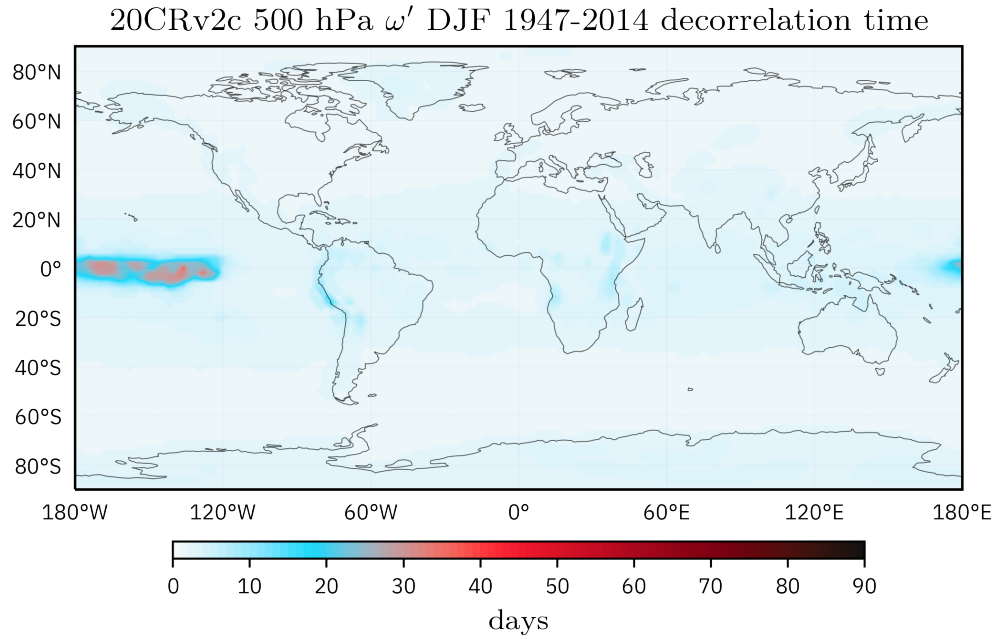


Figure 7.34: Decorrelation time scale in days computed using the $1/e$ autocorrelation threshold for the standardized DJF 500 hPa 20CRv2c vertical velocity anomalies, 1947-2014.

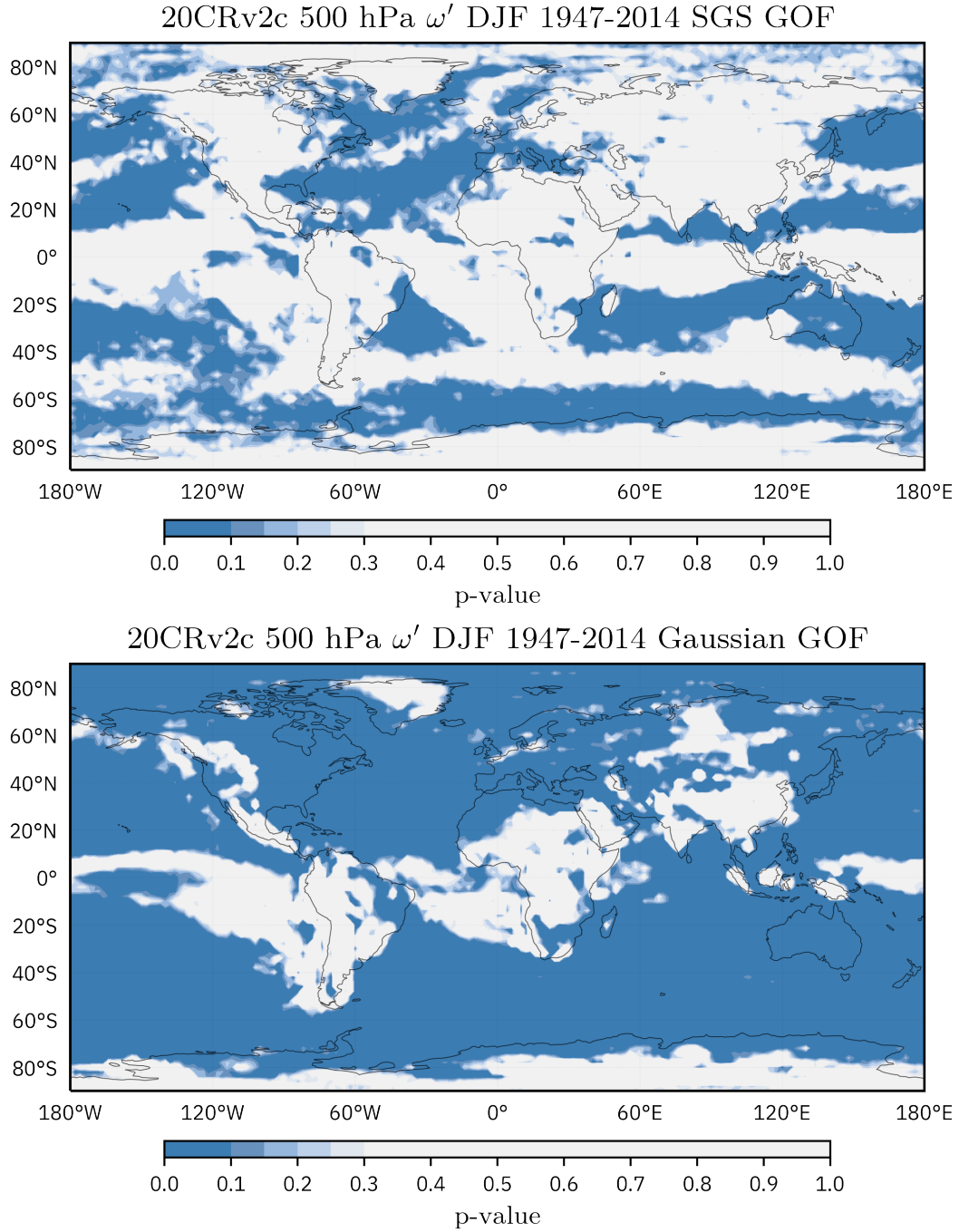


Figure 7.35: p-values for the estimated SGS (top) and standard normal (bottom) goodness-of-fit (GOF) test computed using 100 bootstrap iterations for all gridded time series of standardized DJF 500 hPa 20CRv2c vertical velocity anomalies, 1947-2014. Gray points correspond to a plausible SGS fit, while points gridded blue, especially those with values less than 0.1, correspond to an implausible SGS fit.

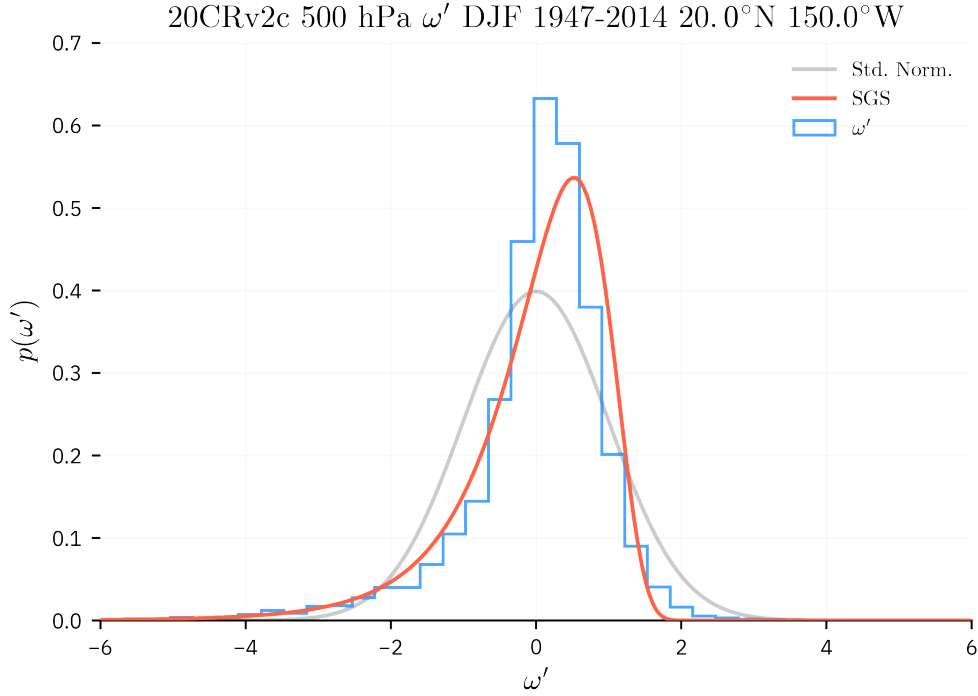


Figure 7.36: SGS pdf (red) and stepwise histograms (blue) of the DJF 500 hPa 20CRv2c vertical velocity anomalies (1947-2014) located at 20°N, 150°W. The standard normal distribution (gray) is provided for reference.

moments to estimate the SGS parameters, has difficulties in producing a plausible fit for vertical velocity anomalies in regions of extreme non-Gaussianity.

Finally, the likelihood ratio test is computed comparing the estimated SGS distribution to that of the standard normal distribution, with the results shown in Figure 7.39. There are few regions of near-Gaussianity in the 500 hPa vertical velocity field, so the estimated SGS distribution is overwhelmingly favored over the standard normal distribution. However, the regions in Figure 7.39 where the estimated SGS distribution is not favored are instead regions of extreme non-Gaussianity, particularly regions of very large kurtosis. This suggests that the SGS distribution has an upper limit on the kurtosis that it can represent, so much so that it fails to beat the standard normal distribution in the likelihood ratio test.

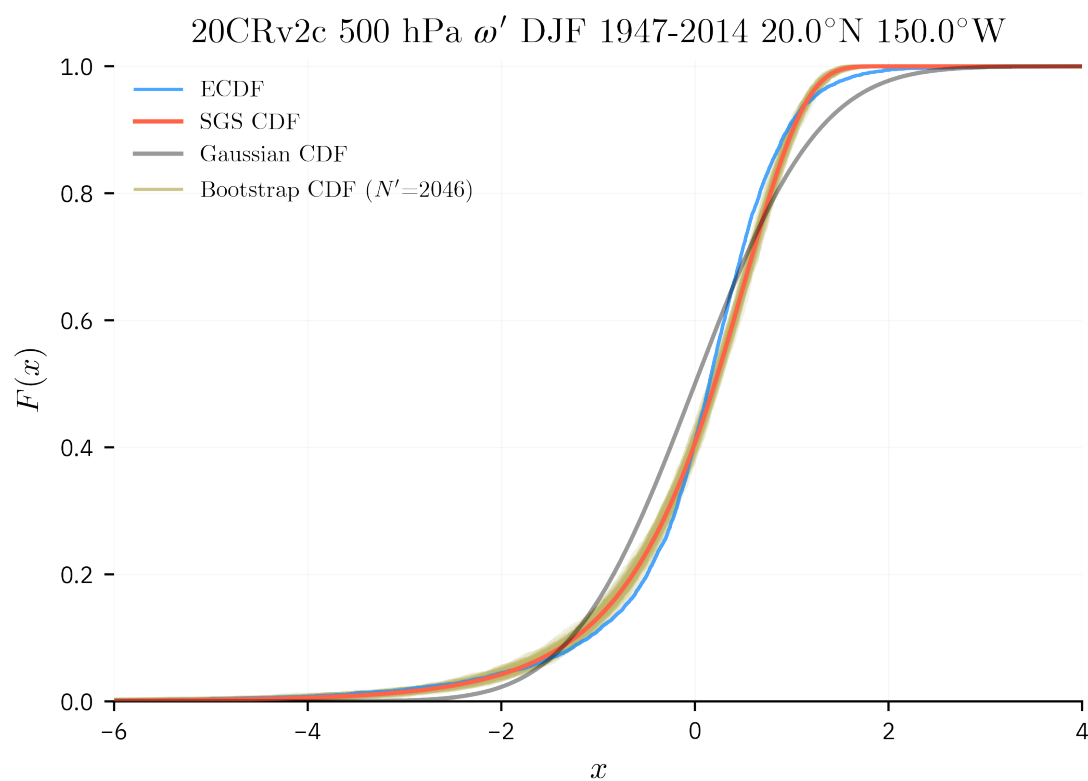


Figure 7.37: Bootstrap CDFs (olive) generated during the standard normal goodness-of-fit (GOF) test for DJF 500 hPa 20CRv2c vertical velocity anomalies (1947-2014) located at 20°N, 150°W, with an equivalent sample size (N') of 2046. The estimated SGS CDF (red) is compared with the ECDF of the time series of anomalies (blue). The standard normal CDF (gray) is provided for reference.

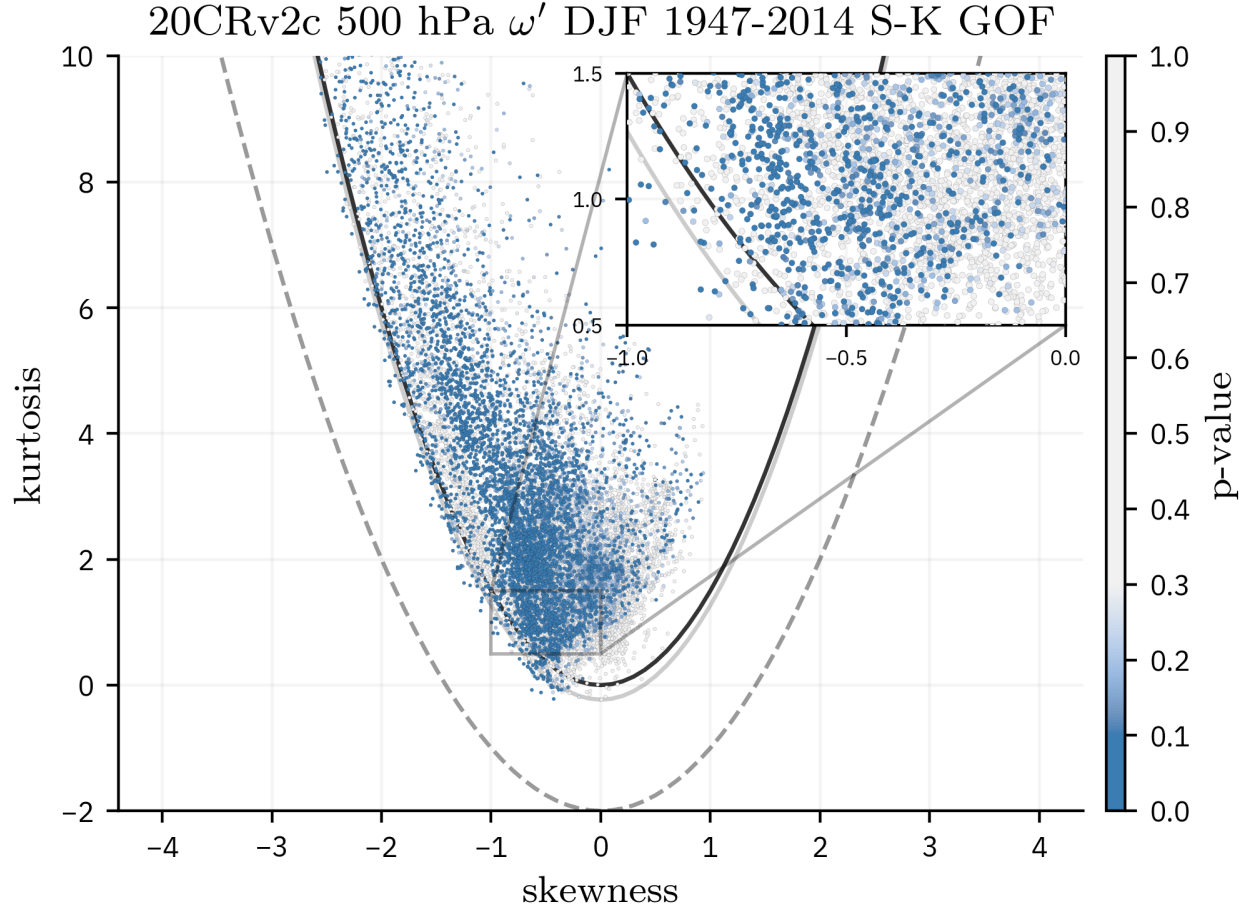


Figure 7.38: Skewness-kurtosis scatter plot showing the goodness-of-fit (GOF) p-value relative to the sample skewness and kurtosis computed for each time series of 500 hPa 20CRv2c vertical velocity anomalies, 1947-2014. The solid black parabola indicates the skewness-kurtosis relationship $kurt \geq \left(\frac{3}{2}\right) skew^2$, where the light gray solid parabola is $kurt \geq \left(\frac{3}{2}\right) skew^2 - r$ where r is empirically fit to the data. The dotted gray line represents the $kurt = skew^2 - 2$ constraint.

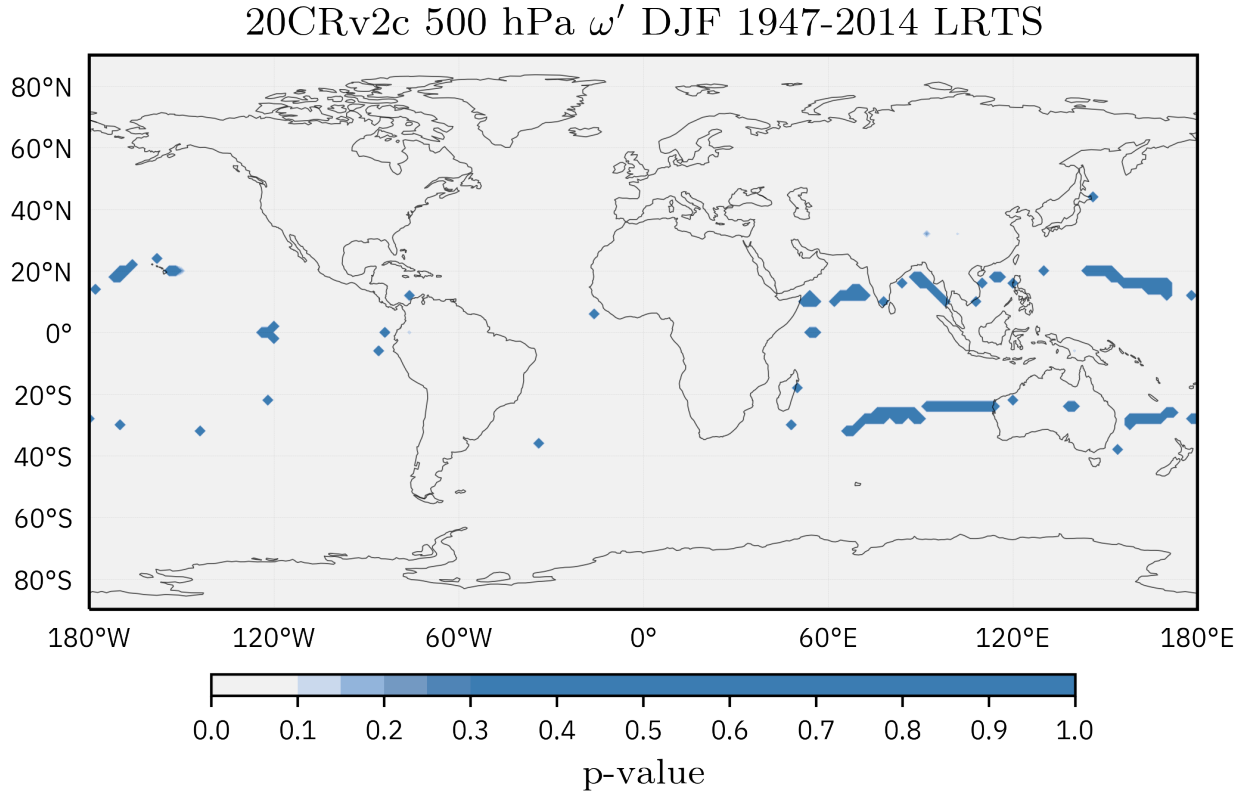


Figure 7.39: p-values of the likelihood ratio test statistic (LRTS) comparison between the estimated SGS and standard normal distributions for all gridded time series of standardized DJF 500 hPa 20CRv2c vertical velocity anomalies, 1947-2014. Gray points correspond to grid points where the SGS distribution is statistically preferred to the standard normal distribution. Points gridded blue, especially those with values much greater than 0.1, correspond to a statistically insignificant preference for the SGS fit or a preference for the standard normal distribution.

CHAPTER 8

A COMPARISON OF REANALYSIS ERAS

The evaluation methodology described in Chapter 6 and applied globally to the 20CRv2c dataset in Chapter 7 is used to make a comparison of the SGS fit from DJF 1947-2014 (6137 days, designated here as the “modern” era) to that of DJF 1880-1947 (6136 days, the “historical” era) for the 950 hPa air temperature, 500 hPa geopotential height and 300 hPa relative vorticity anomalies. Each variable has the annual cycle removed in computing the anomalies, which are then standardized according to the standard deviation of the era. This comparison will show whether the SGS distribution changes significantly across the two periods, or whether the estimated SGS distribution from the entire 1880-2014 period is sufficient to describe the atmospheric non-Gaussian statistics.

8.1 Changes in non-Gaussianity from 1880-1947 to 1947-2014

To begin the comparison of the SGS distributions fit to time series of reanalysis data between the modern (December 1947- February 2014) and historical era (December 1880 - February 1947), the changes in skewness and kurtosis are compared from the different eras. First, Figure 8.1 shows the difference in the 950 hPa air temperature skewness and kurtosis between the historical and modern eras, where the historical moment is subtracted from the modern moment. Positive locations on the map indicate that the moment has increased in magnitude, while negative locations on the map indicate a decrease in moment magnitude. For example, the kurtosis plotted in Figure 8.1 is the kurtosis of DJF 1947-2014 subtracted by the kurtosis of DJF 1880-1947. Figure 8.2 and Figure 8.3 give the difference in the skewness and kurtosis from DJF 1880-1947 to 1947-2014 of the 500 hPa geopotential height anomalies and 300 hPa relative vorticity anomalies, respectively.

Figure 8.1, Figure 8.2 and Figure 8.3 all show a general pattern of decreasing kurtosis in the polar regions, especially near Antarctica. The changes in skewness are relatively uniform and of a smaller magnitude than those of the kurtosis, which can be regionally large. Marked changes include a positive shift in skewness and kurtosis off the western coast of South America in the 950 hPa air temperature anomalies, a hemispherical increase in the magnitude of the 300 hPa relative vorticity

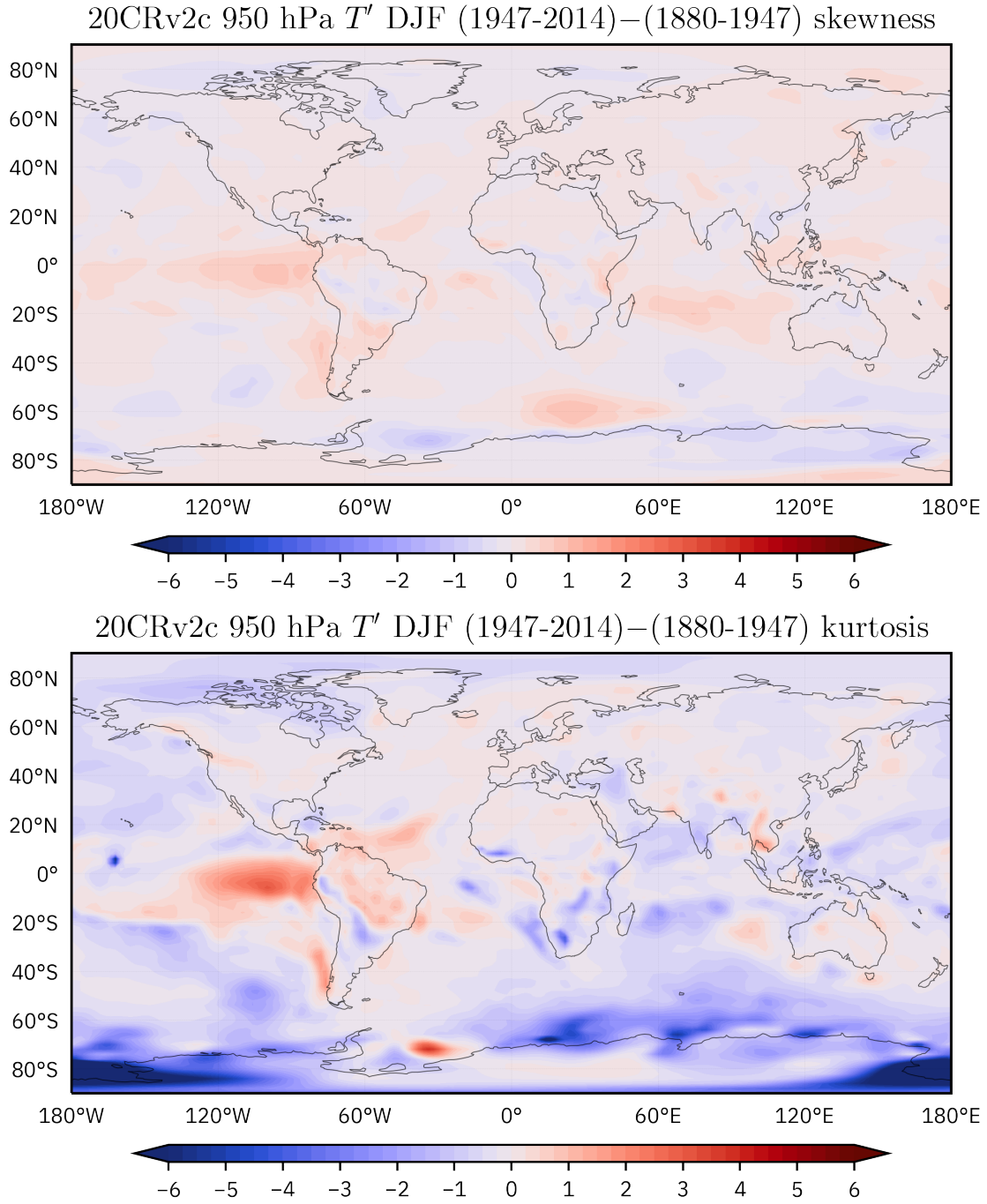


Figure 8.1: Maps of the difference in skewness and kurtosis of standardized 20CRv2c DJF 950 hPa air temperature anomalies from the historical era (1880-1947) to the modern era (1947-2014).

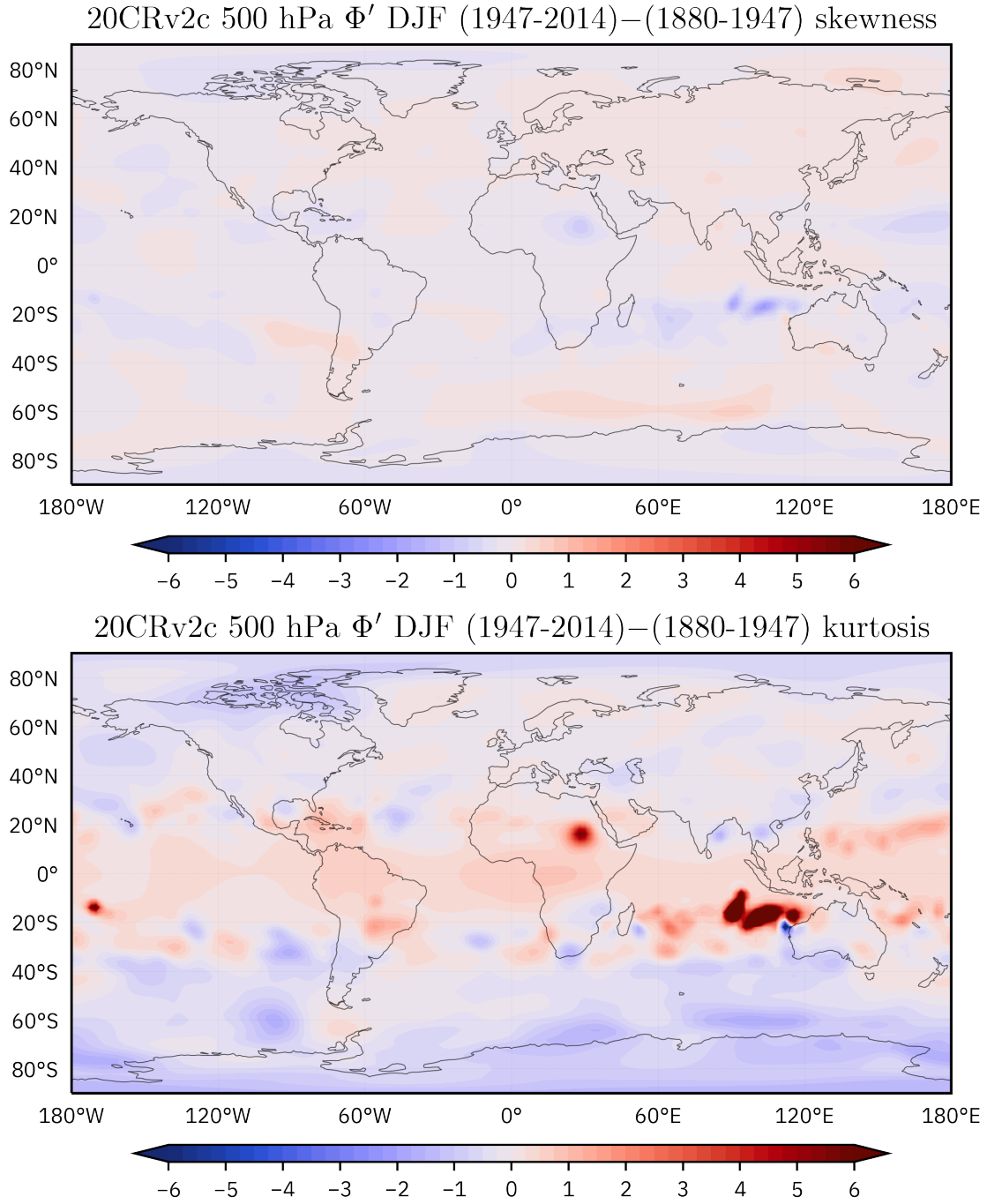


Figure 8.2: Maps of the difference in skewness and kurtosis of standardized 20CRv2c DJF 500 hPa geopotential height anomalies from the historical era (1880-1947) to the modern era (1947-2014).

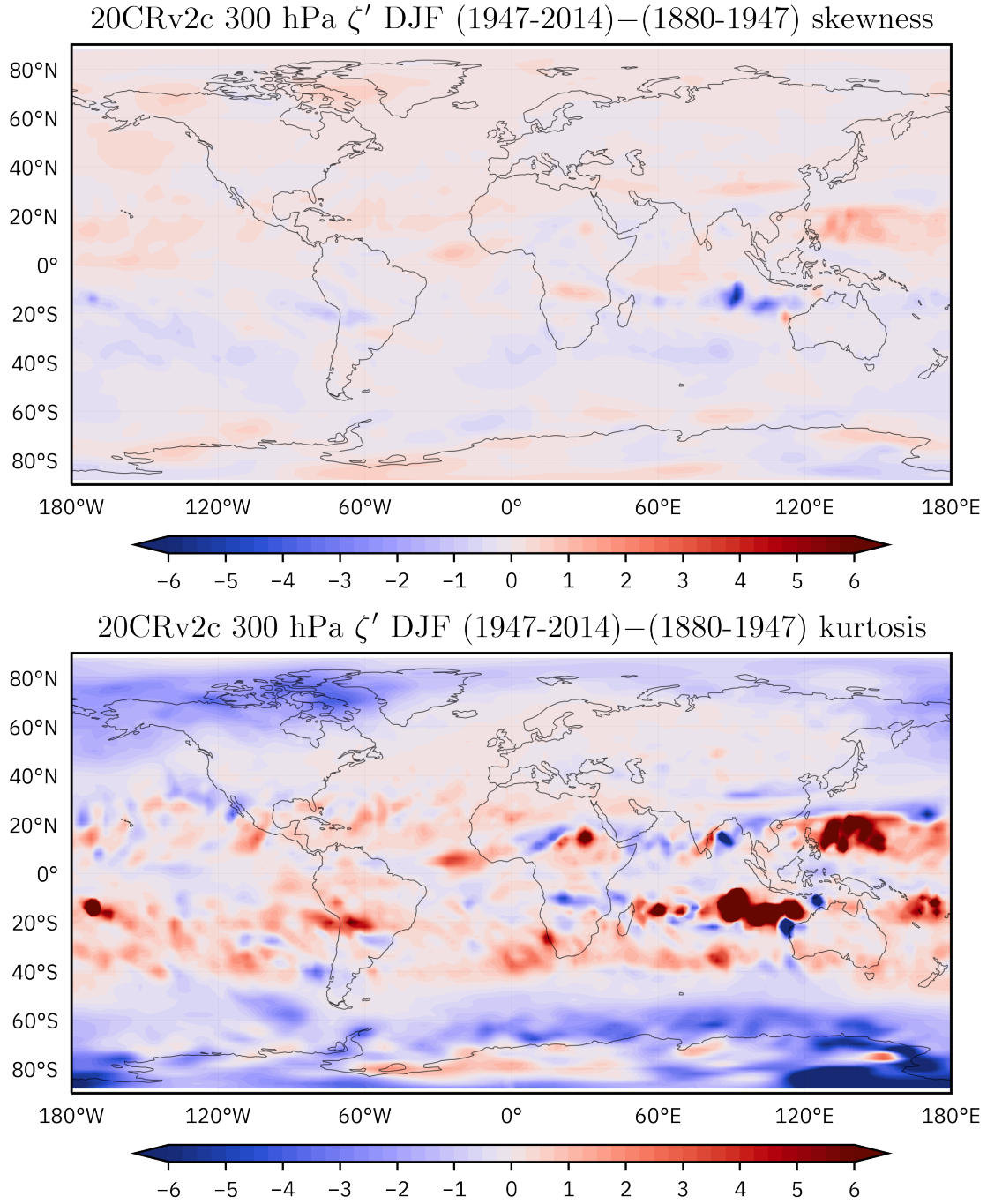


Figure 8.3: Maps of the difference in skewness and kurtosis of standardized 20CRv2c DJF 300 hPa relative vorticity anomalies from the historical era (1880-1947) to the modern era (1947-2014).

skewness and a complex tropical/subtropical banding structure of regional kurtosis increases found in the 300 hPa relative vorticity differences. The 500 hPa geopotential height skewness differences are relatively small, with pockets of substantial kurtosis increases in the tropics.

It is important to note that because the current methodology of computing the SGS distribution is fitting SGS parameters to the sample time series via method of moments (see Section 4.2 and Section 5.3), the changes in the skewness and kurtosis shown in Figures 8.1, 8.2 and 8.3 will drive any changes in the SGS parameters and thus dictate the SGS pdf shape. In other words, as the atmosphere becomes more or less non-Gaussian, the SGS pdfs will diverge more or less from Gaussianity. However, changes in the sample moments do not necessarily directly translate to a change in the SGS distribution. For example, historical and modern era time series with negative sample kurtosis can both violate SGS model constraints the degree that the parameter E will be equivalent for both SGS distributions.

8.2 Changes in the SGS distribution from 1880-1947 to 1947-2014

The two points of 950 hPa air temperature anomalies used in Chapter 6 to describe the statistical evaluation methodology are invoked again in order to inspect how the SGS distribution evolves from the historical era to the modern era. Point A (East Hawaii), located east of Hawaii, is in a region of near-zero skewness and kurtosis. Point B (Vancouver) is located near Vancouver in a region of negative skewness and positive kurtosis. Figure 8.4 shows the estimated SGS pdfs for both eras for the two points, alongside the stepwise histogram of each respective time series of anomalies.

The near-Gaussian East Hawaii time series of anomalies shown in Figure 8.4 sees a contraction to the standard normal distribution due to a decrease in the kurtosis and a slight increase in negative skewness from the historical to modern eras. The non-Gaussian time series of anomalies at Vancouver sees nearly no change, save for a very slight increase in kurtosis. Ultimately from 1880-1947 to 1947-2014, for these two example points, the estimated SGS pdf changes very little, though the near-Gaussian pdf converges slightly to the standard normal distribution while the non-Gaussian pdf diverges further away from the standard normal distribution.

The estimated SGS distributions for both eras at East Hawaii and Vancouver may be compared by computing the SGS CDF of the modern era, then performing a bootstrap goodness-of-fit test in a similar fashion to that of Section 6.3. First, the SGS distributions of the modern and historical

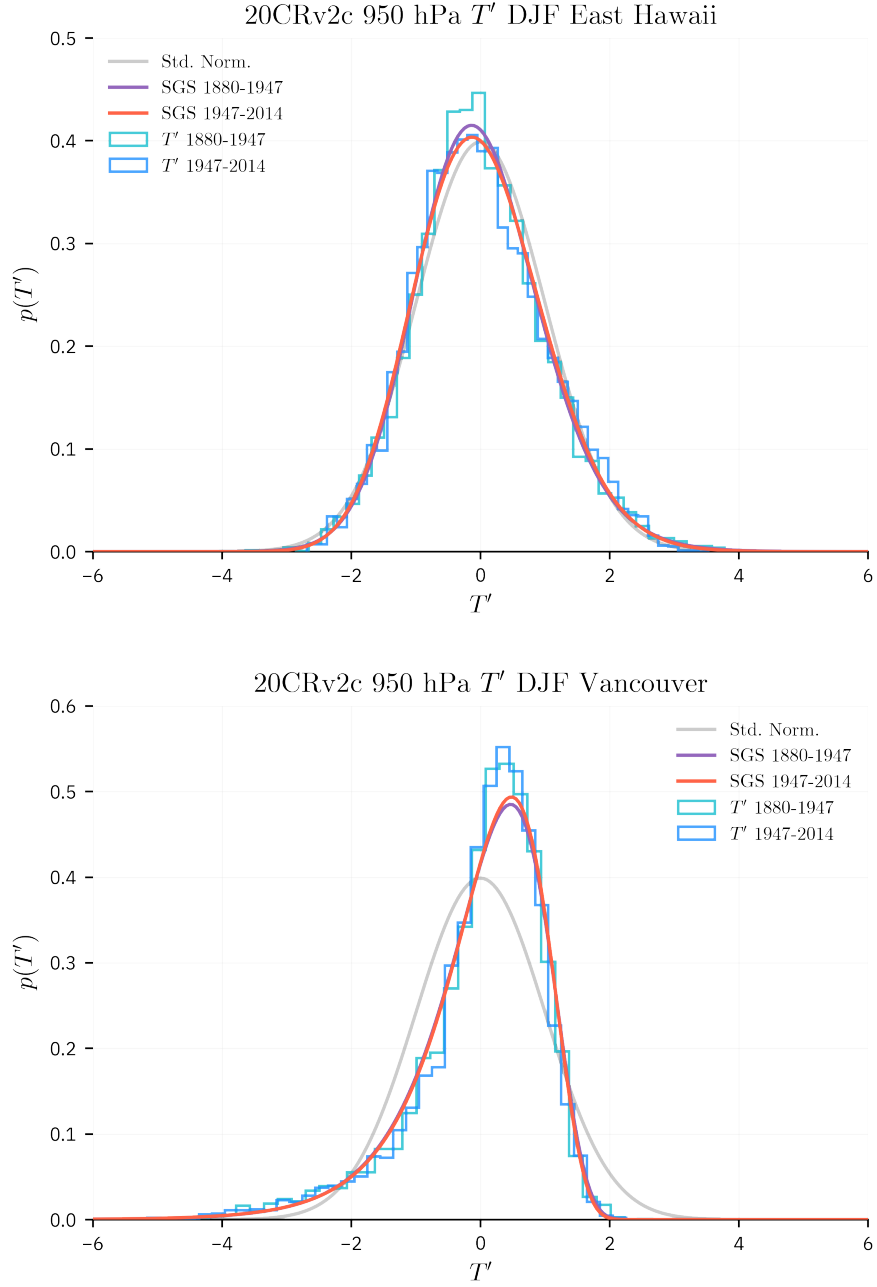


Figure 8.4: SGS pdfs (red, purple) and stepwise histograms (blue, cyan) of the 20CRv2c DJF 950 hPa air temperature anomalies from the historical (1880-1947) and modern (1947-2014) eras at near-Gaussian East Hawaii (top) and non-Gaussian Vancouver (bottom), showing little change between the two time periods. The standard normal distribution (gray) is provided for reference.

eras are calculated. A KS statistic of the modern SGS distribution is then computed using the CDF of the historical era as the reference distribution. Next, pseudorandom datasets drawn from the historical SGS distribution are produced and KS statistics are calculated with respect to the historical CDF. The KS statistics computed from the pseudorandom datasets form a confidence envelope, against which the KS statistic of modern SGS CDF is compared.

Figure 8.5 illustrates this goodness-of-fit comparison of the modern SGS distribution with the pseudorandom ECDFs generated from the historical SGS distribution at East Hawaii and Vancouver, while Table 8.1 shows the results of the goodness-of-fit comparison of eras for both East Hawaii and Vancouver from a trial of 400 pseudorandom datasets.

Table 8.1: The KS statistic and p-value computed via goodness-of-fit test for the near-Gaussian (East Hawaii) and non-Gaussian (Vancouver) DJF 950 hPa air temperature anomaly time series, comparing the estimated SGS distributions of DJF 1947-2014 to that of DJF 1880-1947.

Point	Location	KS Statistic	p-value
A	East Hawaii	0.006	0.98
B	Vancouver	0.004	0.99

To distinguish if the modern SGS distribution is different than the historical SGS distribution, the modern KS statistic is compared against all 400 pseudorandom KS statistics. If the modern SGS KS statistic is less than a KS statistic from a pseudorandom dataset, the modern SGS distribution is assumed to be a better fit to the historical data with respect to the pseudorandom dataset. This KS statistic comparison is performed over all the pseudorandom datasets, where the p-value is computed. Here, the p-value is the ratio of instances where the modern KS statistic is less than the pseudorandom KS statistic relative to the number of pseudorandom datasets. This means that as the p-value approaches zero, the modern SGS distribution is plausibly different than the historical SGS distribution. Otherwise, if the p-value is large and approaches one, no sufficient statistical difference can be determined.

A caveat of applying the goodness-of-fit methodology in this way is that the KS statistic computed between the modern SGS distribution and the reference historical SGS distribution is a direct comparison between continuous distributions obtained from two theoretical CDFs, not em-

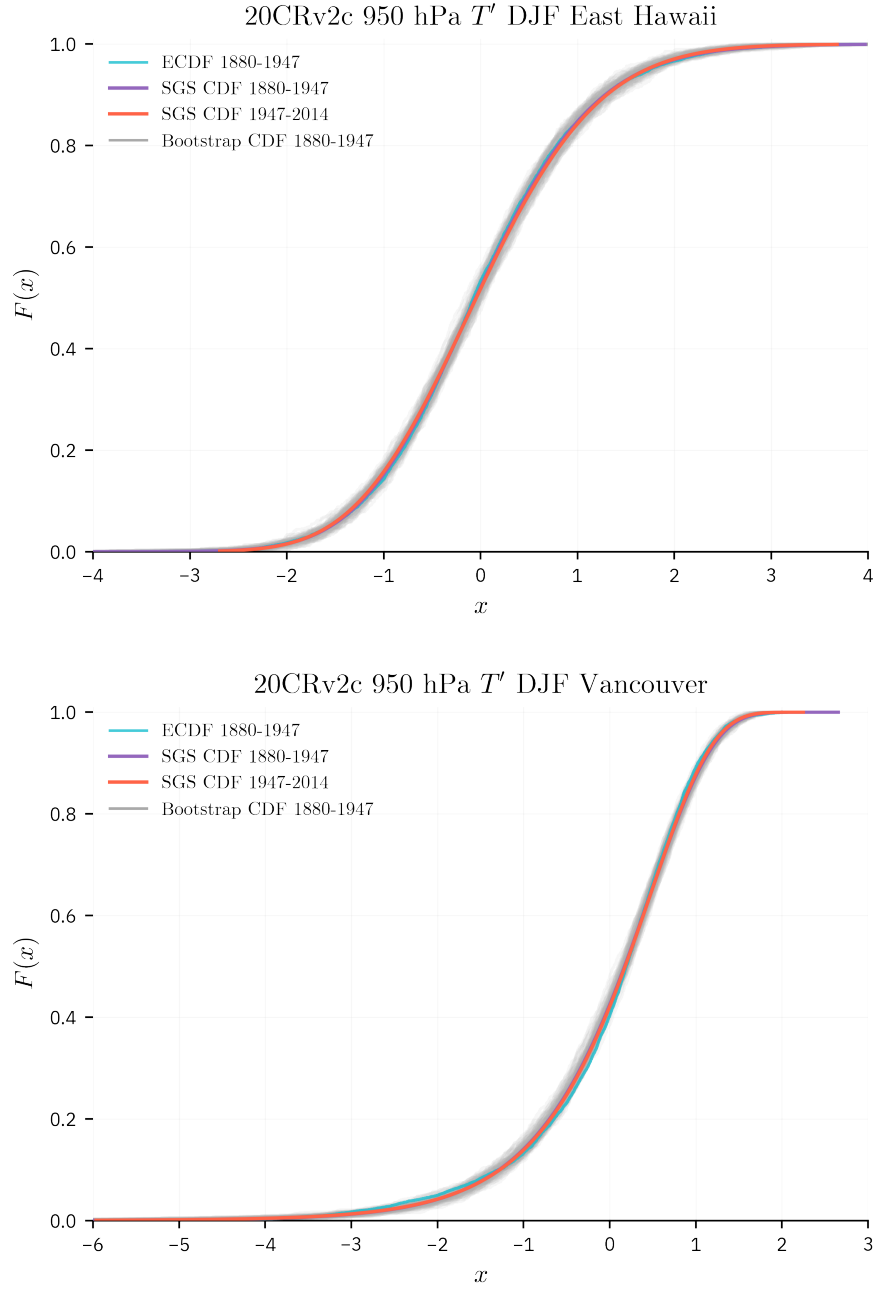


Figure 8.5: ECDFs (blue, purple) of the DJF 950 hPa 20CRv2c air temperature anomalies and estimated SGS CDFs (red, purple) from the historical (1880-1947) and modern (1947-2014) eras at near-Gaussian East Hawaii (top) and Vancouver (bottom). A confidence envelope created with ECDFs obtained from bootstrapping 100 pseudorandom time series is included for comparison.

pirical CDFs. This means that the KS statistic for the modern SGS distribution may be biased lower relative to a comparison of ECDFs. However, as the goal is to compare the shift in the shape of the theoretical distributions, creating an ECDF from the reanalysis time series would not be an appropriate comparison. The era comparison undertaken here is not whether the underlying time series of data changes between eras, but whether the modern SGS distribution is statistically different than the historical SGS distribution.

From Figure 8.5 and Table 8.1, it is clear that for the 950 hPa air temperature anomalies at East Hawaii and Vancouver, the SGS distributions are not plausibly different from each other, as the modern SGS distributions at both the near-Gaussian East Hawaii and non-Gaussian Vancouver are not statistically different than the historical SGS distribution in a majority of cases. For these two locations, the modern SGS distribution is well within the sampling uncertainty of the historical SGS distribution.

Of course, the two 950 hPa air temperature example locations may not be indicative of the overall change in large scale SGS distributions between the two eras. To take a global view of shifts in the SGS distribution, we apply the goodness-of-fit methodology described above to global reanalysis datasets of 950 hPa air temperature anomalies, 500 hPa geopotential height anomalies and 300 hPa relative vorticity anomalies.

Figure 8.6 depicts the goodness-of-fit test comparing SGS distributions fit from 1880-1947 to that of 1947-2014, performed globally for the 950 hPa air temperature anomalies, 500 hPa geopotential height anomalies and 300 hPa relative vorticity anomalies. The grey areas indicate p-values less than 0.1, following the prior p-value threshold set in Chapter 7, which indicate that the SGS distribution of 1947-2014 has plausibly changed relative to the historical era (1880-1947). Blue areas indicate where p-values are greater than 0.1, indicating that the modern SGS distribution was not plausibly different than the historical SGS distribution. The maps of Figure 8.6 for all three 20CRv2c variables generally indicate that the modern SGS distribution is not statistically different from the SGS distribution of the historical era. In line with the method of moments used to calculate the SGS distribution (see 4.2), most locations that indicate statistically plausible changes from the historical SGS distribution to the modern SGS distribution are found in areas where the skewness and kurtosis changed markedly between the two eras. This is true for the 300 hPa relative vorticity anomalies, as gray regions of plausible differences in the SGS distribution are typically confined to

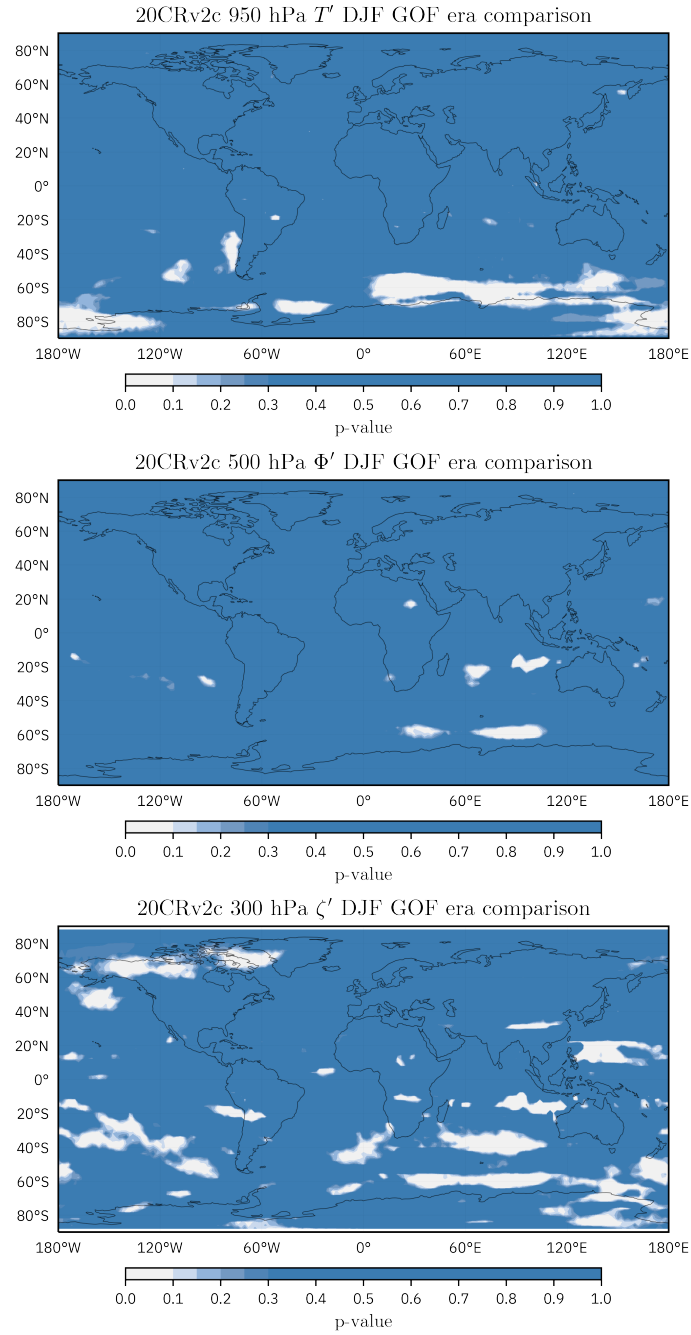


Figure 8.6: p-values obtained from the goodness-of-fit (GOF) test comparing changes in SGS distributions in the modern era (1947-2014) with the historical (1880-1947) era. Gray points correspond to a plausible difference in the SGS distribution between eras, while blue points signify SGS distributions that are not statistically different across the two eras.

locations where the kurtosis (and, at certain locations, skewness) is largely different between the eras, demonstrated in Figure 8.3.

Finally, the “bias” imposed by long decorrelation time scales on the KS statistic, discussed in Chapter 7, is still relevant in interpreting the results of Figure 8.6. Long decorrelation times for points in the tropics, especially for the 950 hPa air temperature (see Figure 7.3) and 500 hPa geopotential height (7.10) differences, result in pseudorandom ECDFs with small equivalent sample sizes. The shorter pseudorandom time series can produce larger KS statistic fluctuations (see Figure 6.6), which raises the KS statistic threshold that the modern SGS distribution must overcome to be classified as different than the historical distribution.

The likelihood ratio test statistic from Section 6.6 can also be invoked to formulate a model selection experiment, where the sum of the likelihood functions over the historical (1880-1947) and modern (1947-2014) eras is tested against the entire reanalysis time period (1880-2014). The likelihood ratio test statistic is written in Equation (8.1),

$$LR = 2((\ell_{\text{hist}} + \ell_{\text{mod}}) - \ell_{\text{full}}), \quad (8.1)$$

where ℓ_{hist} corresponds to the log-likelihood of the SGS distribution from the historical era, ℓ_{mod} corresponds to the log-likelihood of the SGS distribution from the modern era and ℓ_{full} corresponds to the log-likelihood of the SGS distribution from the entire tested reanalysis era. As before, LR corresponds to the likelihood ratio test statistic. Equation (8.1) will test whether the extra parameters obtained through the two SGS distributions are required in order to better represent the non-Gaussian statistics of the full reanalysis era. Here, the log-likelihood for each grid point is computed individually, given by

$$\ell(E, b, g | X) = \sum_{i=1}^N p(x_i, E, b, g), \quad (8.2)$$

where N corresponds to the number of days in the respective reanalysis era. After LR is computed, the χ^2 distribution of Equation (6.5) may again be used to compute a p-value that determines whether the preference of the model selection is due to statistical fluctuations i.e., if the p-value is large or if the selected distribution is strongly preferred (in which case the p-value is small).

Applying the likelihood ratio test statistic of Equation (8.1) to each of the 950 hPa air temperature, 500 hPa geopotential height and 300 hPa relative vorticity anomalies and computing the

p-value using Equation (6.5) with three degrees of freedom (the balance of the free parameters of the combined modern and historical log-likelihoods) yields the respective maps of p-values in Figure 8.7.

The likelihood ratio test statistic of Figure 8.7 has fewer overall grid points that have plausibly different SGS distributions between the two eras compared with Figure 8.6. However, Figure 8.7 conveys the same general message as the goodness-of-fit test, namely that the SGS distribution is represented just as well from the entire reanalysis time period (1880-2014) as it is from adjusting the fit from the historical to the modern SGS distributions. There are a few regions in Figure 8.7 with p-values less than 0.1, which signify an improvement in the representation of the data by fitting an SGS distribution to two eras. However, these regions are generally limited to swaths in the equatorial Pacific Ocean and isolated pockets weighted toward the southern hemisphere.

Figure 8.6 and Figure 8.7 suggests that, at least in the three variables tested here, there is not an improvement to be made for the SGS distributions in the large scale by dividing the reanalysis into two eras. While it is possible that improved results might be obtained by breaking the time domain up into three or more eras, each subsequent division in the time series of anomalies shortens all eras, which eventually leads to statistical results of reduced power.

The lack of evidence suggesting a plausible shift in the shape of SGS distribution for either individual pdfs of air temperature anomalies (as in Figure 8.4) or the majority of geopotential height, air temperature and relative vorticity grid points (given in Figures 8.6 and 8.7) implies that there is no significant mean shift occurring between the two eras themselves. Sardeshmukh et al. (2015) write that, for small climate shifts, SGS parameters E and b may remain constant, but parameter g absorbs the mean external forcing. The change in parameter g leads to changes in the variance, skewness and kurtosis of the SGS distribution, via Equation (4.8). In other words, any mean shift requires a change in the overall shape of the SGS distribution. Conversely, no change in the mean implies no change in the SGS distribution (Sardeshmukh et al. (2015)). Further study is needed to assess whether a significant change in the mean at the spatiotemporal resolution of a reanalysis dataset like 20CRv2c can be statistically detected over (seasonal) eras spanning 67 years (or less). Likewise, empirical analysis is needed to investigate whether a detected mean shift corresponds to a change in the shape of the SGS pdf.

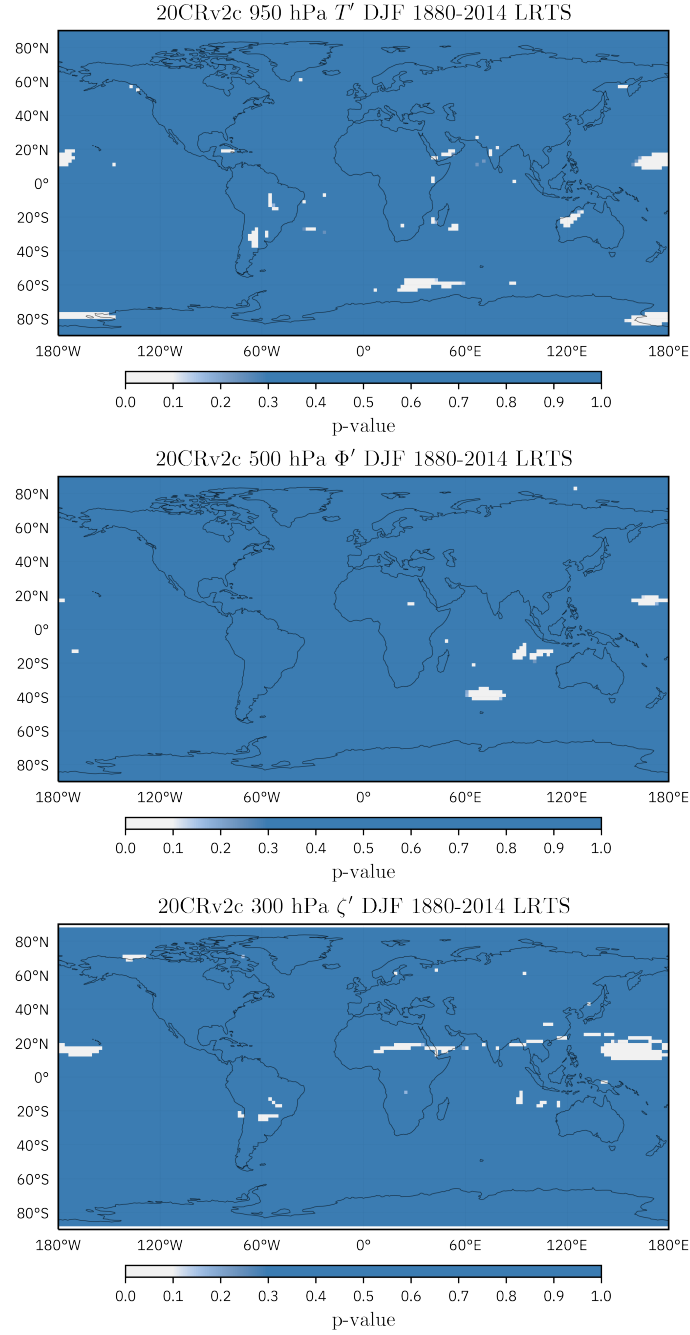


Figure 8.7: p-values obtained from the likelihood ratio test statistic (LRTS) comparing changes in SGS distributions in the modern era (1947-2014) with the historical (1880-1947) era. Gray points correspond to a plausible difference in the SGS distribution between eras, while blue points signify SGS distributions that are not statistically different across the two eras.

CHAPTER 9

CONCLUDING DISCUSSION

9.1 Research overview and conclusions

Despite their infrequency, extreme climate events inflict a disproportionately large impact on lives, property and infrastructure. The statistical study of extreme climate events depends on an accurate representation of non-Gaussian pdfs, as the tails of the pdf are the regions where extreme events occur. However, understanding of the full pdf is needed in order to statistically represent the full climate system dynamics.

The non-Gaussian statistics of daily climate data can be illuminated by calculating statistical moments, such as skewness and kurtosis, which describe the shape of a pdf. The non-Gaussianity of several atmospheric variables from the NOAA-CIRES-DOE Twentieth Century Reanalysis Project version 2c dataset (20CRv2c) was depicted in Chapter 2 with maps of skewness and kurtosis. Unfortunately, the physical mechanisms that drive large-scale patterns of non-Gaussianity are not yet well understood.

Discussed in Chapter 3, stochastic models have been put forward as a novel method to study the non-Gaussian statistics of climate data. In particular, Hasselmann (1976) used stochastic differential equations to approximate climate as a slowly decorrelating process and weather as a rapidly decorrelating process, where climate is forced by weather fluctuations approximated by white noise. However, while Hasselmann’s model produces the familiar red-noise climate spectrum, it is unable to reproduce the non-Gaussian statistics shown in observations. In order to account for the observed non-Gaussian statistics in climate data, Sura and Sardeshmukh (2008) and Sardeshmukh and Sura (2009) introduced a stochastic climate model featuring correlated additive and multiplicative noise. This stochastic model was able to reproduce observed non-Gaussian skewness and kurtosis patterns as well as the red climate spectrum. Chapter 4 discusses results from the CAM noise model, including the stochastically generated skewed (SGS) distribution and a 1-D Markov process model from which SGS-equivalent time series may be produced. Sardeshmukh et al. (2015) introduced a methodology to estimate the parameters of the SGS distribution from sample climate data.

This research evaluated the effectiveness of the SGS distribution in representing the daily non-Gaussian climate statistics using several atmospheric variables from the 20CRv2c dataset. First, a Julia software package, `SGSDist.jl`, was created to effectively estimate and evaluate SGS distributions. Demonstrated in Chapter 5, the `SGSDist.jl` package is able to estimate SGS parameters from sample time series of standardized anomalies, calculate the SGS pdf, CDF and other statistical quantities, draw random numbers from a given SGS distribution and create a time series via Markov process with statistics informed from a particular SGS distribution. This software package has been publicly released under the MIT license.

Chapter 6 makes use of the `SGSDist.jl` package in constructing the statistical evaluation methodology of the SGS distribution with respect to the 20CRv2c dataset. Two sample points of DJF 1947-2014 950 hPa air temperature anomalies were used. SGS distributions were fit to both a near-Gaussian location east of Hawaii and a location of negative skewness and positive kurtosis near Vancouver. Visual inspection of the SGS pdfs and Q-Q plots at each point were made, but a non-parametric test methodology using the Kolmogorov-Smirnov statistic was constructed following Clauset et al. (2009). A likelihood ratio test statistic was also implemented as a model selection test, where the estimated SGS distribution was compared to the standard normal distribution. Caveats to the tests were also discussed, including care that must be taken in regions where a variable’s decorrelation time scale is large.

These statistical testing methodologies were expanded in Chapter 7 to global gridded time series of persistent wintertime (DJF) standardized anomalies of 500 hPa geopotential height, 950 hPa air temperature, 300 hPa relative vorticity, 950 hPa zonal and meridional wind and 500 hPa vertical velocity from 1947-2014. Maps of the skewness and kurtosis as well as SGS model constraint violations encountered during the empirical fitting of the SGS distribution were provided for each variable. The maps of SGS model violations highlighted the difficulty of the method of moments to estimate SGS parameters from a time series with a negative sample kurtosis without violating model constraints.

Goodness-of-fit tests applied to DJF 1947-2014 950 hPa air temperature anomalies showed statistically plausible fits covering the majority of the globe, save for midlatitude regions of negative kurtosis. The estimated SGS distribution performed well in the tropics, but this was certainly aided by the long decorrelation times there, which influences the effective sample size and the ECDF used

to compute the bootstrap KS statistics. The goodness-of-fit test for the standard normal distribution also performed relatively well, likely owing to the small-to-moderate skewness and kurtosis values covering a large area outside the tropics. The likelihood ratio test statistic, however, determined that the estimated SGS distribution was preferred at nearly every grid point across the globe, save those with near-zero skewness and kurtosis.

The DJF 1947-2014 500 hPa geopotential height anomalies produced similar results, in that the SGS model violations were typically confined to latitudinal bands of negative kurtosis. Like that of the air temperature anomalies, the 500 hPa geopotential anomalies also featured a band of large decorrelation times in the tropics. The goodness-of-fit test for the SGS and standard normal cases showed, in all but a few Northern Hemisphere midlatitude locations, that the two distributions were again plausible fits to each gridded time series of anomalies. However, during application of the likelihood ratio test, the SGS distribution was again preferred in all global grid points but the most near-Gaussian of locations.

While SGS model violations for DJF 1947-2014 300 hPa relative vorticity anomalies were again associated with regions of negative kurtosis, long decorrelation times in the tropics were not found. Like the previous two variables, regions of negative kurtosis contributed to a goodness-of-fit test that yielded implausible SGS fits in a few subtropical and midlatitude bands. However, the larger magnitudes of skewness and kurtosis associated with the 300 hPa relative vorticity anomalies contribute to wide swaths of implausible standard normal estimated fits, especially throughout the Southern Hemisphere. Again, the SGS distribution was preferred by the likelihood ratio test in all global locations, save those with near-zero skewness and kurtosis.

The DJF 1947-2014 950 hPa zonal and meridional wind anomalies both featured skewness and kurtosis maps that were noisier than previous variables, with larger regional variability. This was reflected in the goodness-of-fit maps, where regions of positive kurtosis produced plausible SGS fits, while locations of negative kurtosis corresponded to insignificant SGS fits. More implausibly fit regions were found in the standard normal goodness-of-fit tests, especially in the tropics and subtropics where large skewness and (especially) kurtosis values were located. The likelihood ratio test preferred the SGS distribution fit over the standard normal distribution for both cases.

Finally, a goodness-of-fit analysis was undertaken on the DJF 1947-2014 500 hPa vertical velocity anomalies. The skewness and kurtosis maps of vertical velocity are quite different than previous

variables, with near-uniform negative kurtosis outside the poles and near-global positive kurtosis. Also unlike the previous variables, the largest proportion of SGS model violations were confined to the tropics, where large values of skewness and small values of the SGS parameter E contributed to the SGS parameter violation of $b^2 < 0$. A region in the southeastern Indian Ocean also contained skewness values that are larger than the $\sqrt{32}$ bound required for the kurtosis to exist. The goodness-of-fit results of the SGS distribution were quite mixed, with large patches of implausible SGS fits located over subtropical and midlatitude ocean basins. The standard normal goodness-of-fit test was poor, except for regions where the skewness is near-zero and the positive kurtosis was weakest. While the SGS distribution is overwhelmingly favored by the likelihood ratio test, the locations where the estimated SGS distribution is not favored are regions of extreme non-Gaussianity, particularly regions of very large kurtosis. This suggests a possible upper limit on the kurtosis that can be represented by the SGS distribution.

However, it is important to emphasize that the SGS distribution performs quite well over a range of statistics and variables, despite the constraint violations that often arise e.g., difficulty in estimating SGS parameters from time series with negative sample kurtosis. While the statistical performance in regions of negative kurtosis is unfortunate, these locations are not the areas where extreme events statistically occur. In other words, regions of negative kurtosis have smaller pdf tails, which limit the occurrence of extreme events. The performance of the SGS fit is much improved in regions of near-Gaussianity or locations with positive kurtosis. SGS pdfs with positive kurtosis have substantially larger tails, which correspond to the increased likelihood (probability) of extreme events. Therefore, while the SGS fit is not plausible everywhere, it performs best in regions where extreme events are more prevalent.

Chapter 8 applies the SGS evaluation methodology described in Chapter 6 to the comparison of the SGS fit from two different eras (DJF 1880-1947 and DJF 1947-2014) in the 20CRv2c dataset. First, differences in skewness and kurtosis for the three variables are described, which generally depict decreasing kurtosis in polar regions and relatively uniform skewness fluctuations. Skewness changes from DJF 1880-1947 to 1947-2014 were typically of a smaller magnitude than those of kurtosis, which was regionally large in areas. Comparisons between the historical and modern SGS distributions were made between the two previous 950 hPa air temperature locations at East Hawaii and Vancouver, where visual inspection seemed to indicate little change in the SGS shape.

A goodness-of-fit test at both the East Hawaii and Vancouver locations indicated that the SGS distribution did not statistically change from the historical era to the modern era. The globally estimated SGS fits of the 950 hPa air temperature, 500 hPa geopotential height and 300 hPa relative vorticity anomalies were then computed for each 67 year segment and compared using both a goodness-of-fit test and a modified likelihood ratio test statistic. For all three variables examined, the vast majority of grid points indicated that the SGS distribution was not plausibly different in the modern era relative to the SGS distribution of historical era. Changes in the shape of the SGS distribution between the two reanalysis eras is generally limited to areas where the kurtosis is markedly changed between 1880-1947 and 1947-2014.

This research has shown that the SGS distribution is able to describe the non-Gaussian variability observed in several reanalysis variables. The SGS distribution is dynamically consistent and statistically robust and should be strongly considered by future researchers when modeling the full pdf of most daily averaged climate data.

The ability of the CAM noise model to produce time series via Markov process is useful for the creation of ensembles with particular non-Gaussian statistics. The SGS distribution may be estimated from a sample time series, from which many time series realizations may be drawn. This is useful for obtaining confidence intervals for statistical significance testing or risk modeling applications where many thousands of pdfs are typically required to examine tail probabilities. Such an application is quite approachable, as the `SGSDist.jl` package has numerically implemented the CAM noise Markov process, shown in Section 5.7.

As the CAM noise model only represents the internal variability of the climate system, statistics obtained using the SGS distribution may be compared against climate model output that includes external forcing. For example, SGS distributions obtained from sample reanalysis time series may be compared against output from climate models externally forced by varying amounts of greenhouse gas emissions to examine changes in the pdf shape under the respective emission scenarios. Research using this thinking has been employed by Lopez et al. (2018), which used the SGS distribution to measure shifts in summertime 2 meter temperature anomaly pdfs in climate model projections, which in turn was used to attribute some regional extreme heat wave events to anthropogenic climate change.

9.2 Future work

9.2.1 Evaluating climate mean shifts in reanalysis data

While an assessment of the change in the shape of the SGS distribution over two eras was undertaken in Chapter 8, an evaluation of a statistical climate shift is incomplete without examining the change in the mean. As discussed at the end of Section 8.2, a statistical comparison should be made between locations that experience a mean shift to those that experience a change in the shape of the SGS distribution. Such an empirical comparison is required to investigate the claim made by Sardeshmukh et al. (2015) that the shape of SGS distribution remains unchanged when no mean climate shift occurs. More broadly, research into mean shifts in reanalysis data itself would be welcome, as it is not clear whether a significant change in the mean at the spatiotemporal resolution of the 20CRv2c dataset can be statistically detected over temporal eras spanning less than 70 years, especially in data with serial dependence.

9.2.2 Maximum likelihood estimation of SGS parameters

A limitation of the SGS distribution is the difficulty the method of moments (see Section 4.2) has in estimating SGS parameters from a time series that has a negative sample kurtosis without violating the SGS model constraints. Maximum likelihood estimation (MLE) is another commonly used method to estimate parameters of a distribution, where the log-likelihood of a particular function (often times the pdf) is minimized.

A constrained non-linear MLE problem using the SGS pdf as the likelihood function serves as a starting point. Taking the log of Equation (3.29), yields the log-likelihood function for the SGS distribution, written in Equation (9.1) as the objective function to be maximized:

$$\begin{aligned}\ln(L(\dots)) &= l(E, b, g, x_1, \dots, x_n) \\ &= -\log(N) - \left(1 + \frac{1}{E^2}\right) \log((Ex + g)^2 + b^2) + \frac{2g}{E^2b} \arctan\left(\frac{Ex + g}{b}\right)\end{aligned}\tag{9.1}$$

Here, N is the same normalization constant as written in Equation (3.30).

Preliminary tests in estimating SGS parameters using MLE indicate unstable numerical convergence of the objective function in Equation (9.1). Solving for E , b and g using the method of moments can help serve as the “initial guess” for the MLE optimization. However, model constraints, including those that determine if the kurtosis exists, must still be formulated.

9.2.3 Empirical evidence of SGS power law tails

As given in Equation (4.7), when $|Ex + g| \gg b$, the CAM noise is large and the pdf of the SGS distribution can be written in the form of a power law. A continuous power law distribution is characterized by a probability density $p(x)$ of a positive quantity x written as

$$p(x) = Cx^{-\alpha} , \quad (9.2)$$

where C is a normalization constant and the power law exponent, α , must be greater than zero. Sardeshmukh et al. (2015) write the slope of the SGS power law distribution as $2[1 + (1/E^2)]$, which is symmetric with respect to the positive and negative tails. This itself is worthy of further investigation, as symmetric power law slopes are not typically seen in observations (West (2012)).

As the study of extreme events relies on knowledge of the tail of the pdf, opportunities for further research include:

- Examining reanalysis or model output for evidence of power law tails
- Evaluating whether the SGS distribution can statistically represent power law tails in empirical data

There are several techniques that attempt to fit a power law probability distribution to empirical data. Several of these methods (and their limitations) are reviewed in greater detail by Newman (2005) and Clauset et al. (2009). One such method described involves using linear regression techniques after making a log-log histogram of the data. This appears at first to be a reasonable approach, as taking the logarithm of both sides of Equation (9.3) makes it apparent that any power law distribution will appear as a straight line on a log-log plot. However, both Clauset et al. (2009) and Newman (2005) warn that this approach is not preferred for extracting power law distributions as it tends to introduce noise and sampling errors in general circumstances. Clauset et al. (2009) outlines a procedure to fit power law distributions to empirical datasets using a method of maximum likelihood initially developed by Muniruzzaman (1957) as well as a methodology of Clauset et al. (2009) also includes a goodness-of-fit test for evaluating the fit of the power law distribution. This methodology served as a model in this work for the statistical evaluation of the SGS distribution.

Newman (2005) states that almost no physical probability distributions follow a power law over the entire range of x since $p(x)$ diverges when x approaches zero. Therefore, a lower bound on the range of the power law distribution, x_{\min} , is defined as the value of x where the power law

distribution begins. This means the region of the probability distribution where $x \geq x_{\min}$ will contain the power law, if it exists. By normalizing the probability distribution of Equation (9.2) to unity over the domain of the power law and solving for the normalization constant, the normalized power law probability distribution may be written as

$$p(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha}. \quad (9.3)$$

This expression is used to estimate power law distributions from empirical data by determining a unique power law exponent and the domain of the power law distribution.

Currently, there are two serious limitations to the power law solution of the SGS pdf. First, the SGS power law pdf has no cutoff limit, as the power law pdf solution extends out to infinity. This is problematic, as the large-scale atmospheric variables represented in reanalysis or climate model data have a finite domain. It is possible to employ an exponential cut off power law distribution that sets physical bounds on the power law distribution, but modifications to the CAM noise theory would be required to physically motivate such a step. Works such as Stumpf and Porter (2012) reason that without the foundation of a dynamical model, proper statistical analysis and the support of empirical data, the existence of a power law distribution should not be inferred. However, in the case of the stochastic CAM noise climate model of Sura and Sardeshmukh (2008) and Sardeshmukh and Sura (2009), a theoretical basis for the expectation of power laws in SGS pdf tails exists. As such, an examination of the statistical evidence of power law tails using atmospheric data sets is an important step towards understanding the frequency of extreme events.

REFERENCES

- AghaKouchak, A., D. Easterling, K. Hsu, S. Schubert, and S. Sorooshian, 2012: *Extremes in a changing climate: Detection, analysis and uncertainty*. Springer Science & Business Media, Dordrecht, 426 pp.
- Albeverio, S., V. Jentsch, and H. Kantz, 2006: *Extreme events in nature and society*. Center for Frontier Sciences, Berlin, Heidelberg, 352 pp.
- Bartlett, M. S., 1946: On the Theoretical Specification and Sampling Properties of Autocorrelated Time-Series. *Supplement to the Journal of the Royal Statistical Society*, **8**, 27.
- Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah, 2017: Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, **59**, 65–98.
- Box, G. E. P., G. M. Jenkins, and G. C. Reinsel, 1994: *Time series analysis: Forecasting and control*. 3rd ed. Prentice Hall, Englewood Cliffs, N.J., 598 pp.
- Chou, S.-H., E. Nelkin, J. Ardizzone, R. M. Atlas, and C.-L. Shie, 2003: Surface turbulent heat and momentum fluxes over global oceans based on the Goddard satellite retrievals, version 2 (GSSTF2). *Journal of Climate*, **16**, 3256–3273.
- Clauset, A., C. R. Shalizi, and M. E. J. Newman, 2009: Power-Law Distributions in Empirical Data. *SIAM Review*, **51**, 661–703.
- Compo, G. P. and Coauthors, 2011: The Twentieth Century Reanalysis Project. *Quarterly Journal of the Royal Meteorological Society*, **137**, 1–28.
- Cram, T. A. and Coauthors, 2015: The International Surface Pressure Databank version 2. *Geosci. Data J.*, **2**, 31–46.
- Evans, L. C., 2013: *An introduction to stochastic differential equations*. American Mathematical Society, Providence, Rhode Island, 151 pp.

- Franzke, C. L. E., T. J. O’Kane, J. Berner, P. D. Williams, and V. Lucarini, 2015: Stochastic climate theory and modeling. *Wiley Interdisciplinary Reviews: Climate Change*, **6**, 63–78.
- Gardiner, C. W., 2009: *Stochastic methods: A handbook for the natural and social sciences*. 4th ed. Springer, Berlin, 447 pp.
- Giese, B. S., H. F. Seidel, G. P. Compo, and P. D. Sardeshmukh, 2016: An ensemble of ocean reanalyses for 1815–2013 with sparse observational input. *Journal of Geophysical Research: Oceans*, **121**, 6891–6910.
- Hasselmann, K., 1976: Stochastic climate models Part I. Theory. *Tellus*, **28**, 473–485.
- Hirahara, S., M. Ishii, and Y. Fukuda, 2014: Centennial-Scale Sea Surface Temperature Analysis and Its Uncertainty. *Journal of Climate*, **27**, 57–75.
- Holton, J. R., 2004: *An introduction to dynamic meteorology*. 4th ed. Elsevier Academic Press, Burlington, MA, 535 pp.
- Hoskins, B., and T. Woollings, 2015: Persistent Extratropical Regimes and Climate Extremes. *Current Climate Change Reports*, **1**, 115–124.
- Kalnay, E. and Coauthors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, **77**, 437–471.
- Lopez, H., R. West, S. Dong, G. Goni, B. Kirtman, S.-K. Lee, and R. Atlas, 2018: Early emergence of anthropogenically forced heat waves in the western United States and Great Lakes. *Nature Clim Change*, **8**, 414–420.
- Lovejoy, S., 2013: What Is Climate? *Eos, Transactions American Geophysical Union*, **94**, 1–2.
- Luxford, F., and T. Woollings, 2012: A Simple Kinematic Source of Skewness in Atmospheric Flow Fields. *Journal of the Atmospheric Sciences*, **69**, 578–590.
- Messori, G. and Coauthors, 2018: An Interdisciplinary Approach to the Study of Extreme Weather Events: Large-Scale Atmospheric Controls and Insights from Dynamical Systems Theory and Statistical Mechanics. *Bulletin of the American Meteorological Society*, **99**, ES81–ES85.

- Muniruzzaman, A. N. M., 1957: On Measures of Location and Dispersion and Tests of Hypotheses in a Pare to Population. *Calcutta Statistical Association Bulletin*, **7**, 115–123.
- Nakamura, H., and J. M. Wallace, 1991: Skewness of Low-Frequency Fluctuations in the Tropospheric Circulation during the Northern Hemisphere Winter. *Journal of the Atmospheric Sciences*, **48**, 1441–1448.
- Newman, M. E. J., 2005: Power laws, Pareto distributions and Zipf’s law. *Contemporary Phys.*, **46**, 323–351.
- Perron, M., and P. Sura, 2013: Climatology of Non-Gaussian Atmospheric Statistics. *Journal of Climate*, **26**, 1063–1083.
- Petoukhov, V., A. V. Eliseev, R. Klein, and H. Oesterle, 2008: On statistics of the free-troposphere synoptic component: An evaluation of skewnesses and mixed third-order moments contribution to the synoptic-scale dynamics and fluxes of heat and humidity. *Tellus A*, **60**, 11–31.
- Rennert, K. J., and J. M. Wallace, 2009: Cross-Frequency Coupling, Skewness, and Blocking in the Northern Hemisphere Winter Circulation. *Journal of Climate*, **22**, 5650–5666.
- Reynolds, R. W., 1978: Sea surface temperature anomalies in the North Pacific Ocean. *Tellus*, **30**, 97–103.
- Risken, H., 1996: *The Fokker-Planck equation: Methods of solution and applications*. 2nd ed. Springer-Verlag, New York, 472 pp.
- Romanou, A., W. B. Rossow, and S.-H. Chou, 2006: Decorrelation Scales of High-Resolution Turbulent Fluxes at the Ocean Surface and a Method to Fill in Gaps in Satellite Data Products. *Journal of Climate*, **19**, 3378–3393.
- Ruff, T. W., and J. D. Neelin, 2012: Long tails in regional surface temperature probability distributions with implications for extremes under global warming. *Geophysical Research Letters*, **39**.
- Sardeshmukh, P. D., and P. Sura, 2009: Reconciling Non-Gaussian Climate Statistics with Linear Dynamics. *Journal of Climate*, **22**, 1193–1207.

- , G. P. Compo, and C. Penland, 2015: Need for Caution in Interpreting Extreme Weather Statistics. *Journal of Climate*, **28**, 9166–9187.
- Slivinski, L. C. and Coauthors, 2019: Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Q.J.R. Meteorol. Soc.*, qj.3598.
- Sornette, D., 2006: *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools*. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 552 pp.
- Storch, H. v, and F. W. Zwiers, 2001: *Statistical analysis in climate research*. 1st pbk. ed. (with corrections). Cambridge University Press, Cambridge, UK ; New York, 484 pp.
- Stumpf, M. P. H., and M. A. Porter, 2012: Critical Truths About Power Laws. *Science*, **335**, 665–666.
- Sura, P., 2010: On non-Gaussian SST variability in the Gulf Stream and other strong currents. *Ocean Dynamics*, **60**, 155–170.
- , 2011: A general perspective of extreme events in weather and climate. *Atmospheric Research*, **101**, 1–21.
- , and P. D. Sardeshmukh, 2008: A Global View of Non-Gaussian SST Variability. *Journal of Physical Oceanography*, **38**, 639–647.
- , and M. Perron, 2010: Extreme Events and the General Circulation: Observations and Stochastic Model Dynamics. *Journal of the Atmospheric Sciences*, **67**, 2785–2804.
- , and A. Hannachi, 2015: Perspectives of Non-Gaussianity in Atmospheric Synoptic and Low-Frequency Variability. *Journal of Climate*, **28**, 5091–5114.
- Swanson, K. L., and R. T. Pierrehumbert, 1997: Lower-Tropospheric Heat Transport in the Pacific Storm Track. *Journal of the Atmospheric Sciences*, **54**, 1533–1543.
- Trenberth, K. E., and K. C. Mo, 1985: Blocking in the Southern Hemisphere. *Monthly Weather Review*, **113**, 3–21.
- Uppala, S. M. and Coauthors, 2005: The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, **131**, 2961–3012.

- Walpole, R. E., R. H. Myers, S. L. Myers, and K. Ye, 2017: *Probability & statistics for engineers & scientists*. 9th ed. Pearson, Boston, 791 pp.
- West, R., 2012: Power law behavior of atmospheric variability. Florida State University,.
- White, G. H., 1980: Skewness, Kurtosis and Extreme Values of Northern Hemisphere Geopotential Heights. *Monthly Weather Review*, **108**, 1446–1455.
- Wilkins, J. E., 1944: A note on skewness and kurtosis. *The Annals of Mathematical Statistics*, **15**, 333–335.
- Wilks, D. S., 2006: *Statistical methods in the atmospheric sciences*. 2nd ed. Academic Press, Amsterdam ; Boston, 627 pp.
- Williams, P. D., P. L. Read, and T. W. N. Haine, 2003: Spontaneous generation and impact of inertia-gravity waves in a stratified, two-layer shear flow: Generation and impact of gravity waves. *Geophysical Research Letters*, **30**.
- , ——, and ——, 2004: Stochastic resonance in a nonlinear model of a rotating, stratified shear flow, with a simple stochastic inertia-gravity wave parameterization. *Nonlinear Processes in Geophysics*, **11**, 127–135.
- Øksendal, B. K., 2007: *Stochastic differential equations: An introduction with applications*. 6th ed., corr. 4th print. Springer, Berlin ; New York, 369 pp.

BIOGRAPHICAL SKETCH

Robert West was born on 2 August 1986 in Columbia, Missouri. He grew up in Blacksburg, Virginia and is a graduate of Kipps Elementary School. In the summer of 2009 he obtained his B.S. in Meteorology and Applied Mathematics at Florida State University, where he became a graduate student under the instruction of Dr. Philip Sura. He received his M.S. degree in Meteorology in Spring 2012 and defended his PhD dissertation eight years later.